



UNIVERSITÀ DI TRENTO

Emergent information from low-resolution representations

Methods and applications
from protein simulations
to artificial neural networks

Margherita Mele

University of Trento
Department of Physics
Doctor of Philosophy in Physics

Trento, May 2026



UNIVERSITÀ
DI TRENTO

Emergent information from low-resolution representations

Methods and applications
from protein simulations
to artificial neural networks

Margherita Mele

Ph.D. candidate

Supervisor: Raffaello Potestio

Associate Professor, University of Trento

Committee: Silvia Grigolon

*Research Scientist, CNRS & Sorbonne Université
Paris*

Lorenzo Rovigatti

Associate Professor, Sapienza University of Rome

Luca Tubiana

Associate Professor, University of Trento

University of Trento

Department of Physics

Doctor of Philosophy in Physics

Trento, May 2026

Emergent information from low-resolution representations

Copyright © 2026 - Margherita Mele, University of Trento.

This dissertation is original work, written solely for this purpose, and all the authors whose studies and publications contributed to it have been duly cited. Partial reproduction is allowed with acknowledgment of the author and reference to the degree, academic year, institution—*University of Trento*—and public defense date.



Abstract

We live in an age characterised by informational abundance, in which not only scientific research but also contemporary society in general is shaped by the unprecedented production, availability, and diffusion of data. On the one hand, advances in measurement techniques, simulation methods, and computational power have enabled the description of natural systems at increasingly fine levels of detail and across broader spatial and temporal scales. On the other hand, the massive amount of digital data constitutes the essential fuel that feeds learning systems, making this abundance not only an epistemic resource for science but also a constitutive condition for the development of artificial intelligence.

This thesis addresses this problem within a framework grounded in statistical physics and information theory, in which the configurational space provides the natural setting for the description of complex systems, and reduced representations – namely simplified descriptions of a more detailed object of study – are treated as tools for identifying the system properties most relevant to collective behaviour. From this perspective, the central questions that arise are how the informational content of a low-resolution description should be quantified; under what conditions informative reduced representations can be identified without supervision; and to what extent common organising principles can be recognised across systems of different microscopic nature, such as proteins and artificial neural networks. More generally, the aim of the present work is to clarify how reduced descriptions can provide not only computational simplification, but also conceptual access to the emergent organisation of complex systems.

Keywords: Coarse Graining, Information Theory, Complex Systems

List of Publications

Peer-reviewed journal articles

- **Mele, M.**, Fiorentini, R., Tarenzi, T., Mattiotti, G., Potestio, R. (2026). Determining the optimal structural resolution of proteins through an information-theoretic analysis of their conformational ensemble. *Journal of Chemical Theory and Computation*.
- **Mele, M.**, Menichetti, R., Ingrosso, A., & Potestio, R. (2025). Density of states in neural networks: an in-depth exploration of learning in parameter space. *Transactions on Machine Learning Research*.
- **Mele, M.**, Covino, R., Potestio, R. (2022). Information-theoretical measures identify accurate low-resolution representations of protein configurational space. *Soft Matter*.

Preprints

- **Mele, M.**, Campos Moreno, D., Potestio, R. (2026). The bliss of dimensionality: how an unsupervised criterion identifies optimal low-resolution representations of high-dimensional datasets. *In review*; available as preprint.

In Preparation

- **Mele, M.**, Castagna, A., Menichetti, R., Ingross, A., & Potestio, R. (2026). Unsupervised Information-Theoretic Pruning of Shallow Neural Networks.

Other manuscripts in preparation not part of this thesis

- Rocca, A., **Mele, M.**, Röder, K., Pasquali, S. & Potestio, R. (2026). Critical variable selection to discriminate the interplay between features and metrics in biomolecular simulations.
- Spreti, C., **Mele, M.**, Potestio, R. (2026). Enhancing enhanced Sampling: boosting Wang-Landau microcanonical sampling for off-lattice Soft Matter Systems.

For manuscripts in preparation, titles and author lists are provisional and may change before submission.

Contents

<i>Acronyms</i>	x
1 Introduction	1
Proteins	
2 Theoretical Background	9
2.1 Molecular Dynamics	11
2.1.1 From a Quantum to a Classical Description	11
2.1.2 Empirical Force Fields	11
2.1.3 Integration of the Equations of Motion	13
2.1.4 Statistical Ensembles and Thermodynamic Control	14
2.2 Coarse-Graining: From Multiscale Systems to Reduced Representations	19
2.2.1 Multiscale Structure and the Renormalization Group Perspective	19
2.2.2 Formal Statistical–Mechanical Framework	20
2.2.3 Philosophies of Coarse-Graining	23
2.2.4 Resolution Level and Representational Power	26
2.2.5 The Mapping Problem	27
2.2.6 Coarse-Graining as an Analytical Operator	29
2.3 The Resolution-Relevance Framework	30
2.3.1 Resolution: Complexity and Coding Cost	31
2.3.2 Relevance: Informative Variability in Coding Cost	34
2.3.3 Resolution and Relevance in Practice	35
2.3.4 The Resolution–Relevance Tradeoff	38
2.3.5 Information-Theoretic Bounds	40
2.3.6 Multiscale Relevance	44
3 The Bliss of Dimensionality	47
3.1 Introduction	47
3.2 Materials and Methods	48
3.2.1 Discrete Representations	48
3.2.2 Relevance-Resolution and Supervised Optimality	49

3.2.3	Systems and Datasets	51
3.3	Results	53
3.3.1	Unstructured Synthetic Datasets	54
3.3.2	Structured Synthetic Datasets	55
3.3.3	Gaussian Clones of MNIST	56
3.3.4	Alanine Dipeptide	58
3.4	Conclusions	59
4	Configurational Coarse-Graining	63
4.1	Introduction	63
4.2	Materials and Methods	64
4.2.1	Systems and Simulation Data	65
4.2.2	Diffusion Maps and Covariance Decomposition	67
4.3	Results	68
4.3.1	12-Protein Dataset	68
4.3.2	Adenylate Kinase	70
4.4	Conclusions	74
5	Structural Resolution	77
5.1	Introduction	77
5.2	Materials and Methods	79
5.2.1	Systems and Simulation Data	79
5.2.2	Structural Decimation and Induced Configurational Clustering	80
5.2.3	PROPRE Analysis and Identification of the Optimal Resolution	81
5.2.4	Structural Fluctuation Analysis	82
5.2.5	Mapping Entropy	82
5.3	Results	83
5.3.1	Optimal Number of Sites <i>vs.</i> Number of Protein Residues . .	84
5.3.2	Optimal Number of Sites <i>vs.</i> Position Covariance	87
5.3.3	Optimal Number of Sites <i>vs.</i> Mapping Entropy	88
5.4	Conclusions	90
Networks		
6	Theoretical Background	95
6.1	Feed-Forward Neural Networks and the Learning Problem	97
6.1.1	Feed-Forward Neural Networks	97
6.1.2	Supervised Learning, Training Error, and Generalisation . . .	98
6.1.3	Gradient-Based Optimisation	99
6.1.4	The Teacher-Student Setting	101
6.1.5	Beyond Feed-Forward Architectures	102
6.2	Wang-Landau Sampling	103
6.2.1	Density of States and Thermodynamic Reconstruction	104

6.2.2	Generalised-Ensemble Motivation and the Flat-Histogram Idea	105
6.2.3	The Original Wang-Landau Algorithm	106
6.2.4	Practical Aspects of the Wang-Landau Method	108
6.2.5	Error Saturation in the Original Wang-Landau Scheme	111
6.2.6	The $1/t$ Algorithm	113
6.3	Applications and Broader Scope of Wang-Landau Sampling	115
6.4	Mapping Entropy	117
6.4.1	Relative Entropy Framework	118
6.4.2	Information-Theoretic Interpretation	120
6.4.3	Practical Evaluation of the Mapping Entropy	123
7	Mapping Entropy Multi-Body Decomposition	133
7.1	Introduction	133
7.2	Formal Framework	134
7.3	Results	135
7.3.1	Multi-body Decomposition of the Entropy	135
7.3.2	Multi-body Decomposition of the Mapping Entropy	137
7.3.3	Non-Interacting Model	139
7.3.4	Computational Illustration of the Non-Interacting System	146
7.4	Conclusions	150
8	Neural Network Configuration Space	153
8.1	Introduction	153
8.2	Materials and Methods	154
8.2.1	Neural Network Architectures	154
8.2.2	Datasets	156
8.2.3	Wang-Landau Sampling of the Parameter Space	157
8.3	Results	158
8.3.1	Density of States Analysis of Benchmark Tasks	159
8.3.2	Density of States Analysis of Synthetic Tasks	161
8.4	Conclusions	167
9	Neural Network Dropout Space	171
9.1	Introduction	171
9.2	Materials and Methods	173
9.2.1	Teacher-Student	174
9.2.2	Non-Linear Gaussian Process	176
9.2.3	Wang-Landau Sampling of the Dropout Space	178
9.2.4	Performance Evaluation of Reduced Networks	179
9.3	Results	179
9.3.1	Structure of the Mappings Selected by Mapping Entropy Minimisation	179
9.3.2	Performance of the Reduced Network	184

9.4	Conclusions	187
10	Conclusions	189
Appendices		
A	Optimality Point Identification	195
B	SI: Bliss of Dimensionality	201
B.1	Gaussian clones of MNIST: exclusion of digit 1	201
B.2	Additional structured-synthetic parameter combinations	204
B.3	Additional analyses on the KL criterion	205
C	SI: Configurational Coarse-Graining	209
C.1	Additional analyses for adenylate kinase	209
C.2	Additional validation on the humanised IgG4 monoclonal antibody	212
D	SI: Structural Resolution	215
D.1	Extended fit on the 11-protein dataset	215
D.2	Quantitative comparison with Martini and SIRAH	216
D.3	Robustness and Reproducibility of the PROPRE estimates	218
D.3.1	Robustness with respect to stochastic sampling of the map- ping space	218
D.3.2	Reproducibility across independent simulation replicas . . .	219
E	SI: Mapping Entropy Multi-Body Decomposition	221
E.1	Reference local measure and beyond the uniform degeneracy as- sumption	226
F	SI: Neural Network Configuration Space	229
F.1	Additional details on the datasets	229
F.1.1	Synthetic data	229
F.1.2	Gaussian Clones	230
F.2	Additional details on the Wang-Landau algorithm	231
F.3	Validation and scaling analysis	234
F.3.1	Validation of entropy estimation	234
F.3.2	Scaling analysis of the Wang-Landau algorithm	235
F.4	Gaussian Fit	236
F.5	Fashion-MNIST	238
	<i>Bibliography</i>	243

Acronyms

AI	Artificial Intelligence. (<i>p.</i> 95, 96)
CG	Coarse-Graining. (<i>p.</i> 5, 18–20, 22–24, 26–31, 33, 37–40, 44, 45, 47, 48, 63, 64, 66, 68–70, 73–78, 82–84, 87, 88, 90, 91, 109, 116–123, 125, 127, 129, 137, 141, 143–147, 150, 151, 173, 178, 180, 184, 187, 188, 190, 191, 198, 217, 218)
DoS	Density of States. (<i>p.</i> 154, 156–167, 169, 178, 179, 183, 184, 190, 230, 231, 234, 236–239)
KL	Kullback-Leibler. (<i>p.</i> 48, 49, 51–61, 118–123, 125–127, 129, 134, 139, 201, 205–208)
MB-PMF	Many-body Potential of Mean Force. (<i>p.</i> 22, 23, 118)
MD	Molecular Dynamics. (<i>p.</i> 3, 5, 10, 13, 15–19, 24, 53, 63, 64, 68, 74, 76, 77, 79, 129, 190, 220)
ME	Mapping Entropy. (<i>p.</i> 6, 24, 92, 117, 118, 120–131, 133–135, 137–151, 173, 174, 176–190)
MSR	Multiscale Relevance. (<i>p.</i> 65–70, 72, 74–76)
NLGP	Non-Linear Gaussian Process. (<i>p.</i> 173, 176, 178, 179, 182–187)
RR	Resolution-Relevance. (<i>p.</i> 5, 30, 31, 38, 44, 45, 47–51, 53–61, 64–75, 78, 80, 81, 84, 88, 90, 91, 143, 189, 195, 197, 198, 201–205, 207–213, 215, 216)
TS	Teacher-Student. (<i>p.</i> 101, 102, 157, 173, 174, 176, 179, 180, 182, 187)
WL	Wang-Landau. (<i>p.</i> 5, 6, 103, 104, 106, 108–117, 131, 154, 157, 158, 167–169, 178, 184, 229)

1

Introduction

If a hundred army ants are set loose on a flat surface, they will march in endless circles until they die (Schneirla, 1944). However, when half a million of them gather, they move as one, swarming through the forest in coordinated waves. They build living bridges and shelters and hunt with astonishing efficiency. How does this transformation, from confusion to apparent intelligence, take place? Similar mysteries arise when cells coordinate to form organs, when neurones ignite to create thought, and when the immune system mounts a defence without a commander. Systems like these, where countless simple units combine to generate surprising order and function, are known as complex systems.

The first and most fundamental definition of a complex system is: a set of interconnected components whose collective behaviour cannot be simply explained by the behaviour of each individual element (Anderson, 1972). This simple statement captures the notion that the whole is qualitatively distinct from the sum of its parts. However, this apparently simple definition has long been debated (Ladyman et al., 2013). How many elements are considered to be “many”? At what point does a system cross from being merely complicated to being truly complex? Although a watch has thousands of parts, its behaviour is fully determined by its mechanism. In contrast, the brain of a worm is made up of only a few hundred neurones, yet it exhibits behaviours that far exceed the sum of its components. The boundary has never been sharp, and this lack of consensus has fuelled decades of discussion.

The editors of the *Journal of Physics: Complexity* recently took the opportunity, following the 2021 Nobel recognition of work on climate models and disordered systems, to collectively reflect on the state of the field (Bianconi et al., 2023). Their contributions highlight a diversity of perspectives ranging from physics and biology to social science and engineering, underscoring the inherently interdisciplinary nature of complexity science. Rather than being a disadvantage, this plurality illustrates the wide range of phenomena that fall under the umbrella of complex systems and the continuing importance of studying them. Complexity science is an evolving field that breaks down traditional disciplinary boundaries while offering a common

language with which to address some of the most fundamental scientific questions and urgent social issues.

At the heart of these perspectives lies the idea that complex systems are formed by many interacting elements whose collective behaviour gives rise to *emergent* phenomena. Emergence refers to the appearance of properties or patterns that cannot be deduced from the behaviour of individual components in isolation. A single neuron can only fire or remain silent, but in networks of billions, their coordinated activity gives rise to perception, memory, and thought. In each case, the whole is not simply larger than its parts, but qualitatively different, requiring a new language of description. This highlights a paradox: a complete knowledge of the microscopic elements and their dynamics does not necessarily help us to understand the system and can even become detrimental if it obfuscates the collective principles at play. From the viewpoint of statistical physics, this is precisely what makes complexity compelling. It forces us to confront the limits of reductionism and to ask when a microscopic description is informative and when, instead, a coarse collective perspective is essential (Laughlin et al., 2000).

A second fundamental theme is that complexity often emerges in the interplay between order and randomness. Systems that are too ordered become rigid and predictable; systems that are too random dissolve into noise. Between these extremes lies a fertile regime where local irregularities coexist with large-scale organisation (Langton, 1990). Statistical physics has offered two influential paradigms to capture this balance. The network perspective stresses the importance of heterogeneity in the pattern of connections: a small amount of randomness in the way elements are linked can produce hubs and bottlenecks that drastically affect collective behaviour (Albert et al., 2002; Watts et al., 1998). The spin-glass perspective, in contrast, emphasises randomness in the sign of interactions: mixtures of cooperation and competition generate frustration and give rise to rugged landscapes with many nearly equivalent states (Mézard et al., 1987). Taken together, these viewpoints illustrate a central lesson of complexity science: what defines a complex system is not only the large number of interacting elements, but also the architecture, heterogeneity, and diversity of their interactions, all together shaping the emergent behaviour of the whole.

A third central aspect in the study of complex systems is that many apparently different systems show comparable patterns of collective behaviour. Epidemics, financial markets, ecosystems, and communication networks are built on entirely different substrates and follow distinct microscopic rules, yet they all display similar large-scale features: sudden transitions (Scheffer et al., 2001), cascades of failures (Bak et al., 1987), and a characteristic asymmetry between robustness and fragility – resilient to random perturbations, but extremely sensitive to targeted ones (Albert et al., 2000). This recurrence of patterns across very different domains suggests that complexity is not tied to any specific material or mechanism but instead reflects organisational principles that appear again and again in nature and society.

Understanding these recurring principles – what statistical physics calls universality – has long been one of the core goals of the field (Anderson, 1972; Goldenfeld et al., 1999).

The last aspect common to most complex systems is their hierarchical organisation. Their behaviour cannot be described at a single level but arises from the interaction between processes operating across multiple scales. Molecules combine to form cells, cells assemble in tissues, and tissues give rise to organisms; individuals form communities, communities structure societies, and societies create institutions and cultures (Simon, 2012; Bar-Yam, 2019; Bar-Yam, 2004). Each level introduces new emergent structures and governing principles, while interactions between levels ensure that local processes shape global behaviour, and global organisation, in turn, constrains the local dynamics. This multiscale structure is both a defining source of complexity and the reason why purely reductionist schemes so often fail, since every scale contributes its own degrees of freedom in an interdependent manner.

Each organisational level is regulated by its unique characteristics, timescales, and interactions. To understand a system, it is often necessary to know the specific size at which the phenomena of interest occur. Consider the example of a fluid: at a microscopic level, the stochastic motion of individual particles, collisions, and fluctuations can be explained by the interatomic forces and collisions that govern Molecular Dynamics; at a larger scale, however, the same system can be considered a continuous medium that satisfies hydrodynamic equations. The microscopic description is necessary to explain diffusion and viscosity, but not large-scale structures such as vortices and waves, which can only be captured by the continuum picture where molecular interactions are neglected. The correct description, therefore, depends on the question: properties vary across scales, and no single level provides the full picture. When the characteristic scales of the underlying processes are widely separated, they can generally be solved independently, building useful theories that are valid at each level of granularity. This *separation of scales* is a guiding principle in much of contemporary physics, ranging from thermodynamics to molecular biology and even economics. In complex systems, however, scale boundaries blur, with correlations and feedback coupling micro and macro dynamics in such a way that often no single level of description can be adequate in isolation (Mora et al., 2011).

All the characteristic features discussed so far – emergence, hierarchy, heterogeneity, and multiscale coupling – are ultimately rooted in what is referred to as the configurational space of the system. This space represents the multidimensional set of all microscopic states that the system can, in principle, access. Each point within it corresponds to a complete specification of the system's *fundamental* variables: the positions of atoms in a molecule, the orientations of spins in a magnetic material, or the activity pattern of neurons within a network. In essence, the empirical information available for studying a complex system consists of the ensemble of these mi-

croscopic configurations, which collectively define its configurational space. From the perspective of statistical physics, the configurational space provides the natural framework in which the system's dynamics unfold. As the system evolves, it traces a trajectory across this high-dimensional landscape. The geometry and topology of the space – its basins, barriers, and connecting pathways – govern the system's behaviour: whether it remains confined within certain regions, undergoes abrupt transitions, or reorganises in response to external perturbations (Stillinger et al., 1984; Wales, 2004; Frauenfelder et al., 1991). Exploring the structure of configurational space, therefore, offers a unifying way to connect microscopic interactions with macroscopic phenomena, elucidating how local interaction rules give rise to collective organisation and emergent behaviour.

Classic examples illustrate how analysing configurational space has transformed our understanding of complex systems. In the spin-glass model, exploring the energy landscape of spin orientations revealed the presence of frustration and a multitude of metastable states, explaining phenomena such as memory effects and slow relaxation (Mézard et al., 1987). In molecular systems, studying the conformational landscape of atomic positions uncovered the existence of folding funnels and transition pathways that govern stability and function (Stillinger, 1995; Onuchic et al., 1997; Bryngelson et al., 1987). In network systems, mapping the space of connection weights or activity patterns has shed light on how learning and generalisation arise from the geometry of the loss landscape (Choromanska et al., 2015; Ballard et al., 2017). Across these examples, characterising configurational space exposes the effective degrees of freedom, dominant pathways, and collective mechanisms that are invisible from a purely microscopic perspective.

Despite being conceptually rich, the analysis of configurational spaces poses formidable challenges. Its dimensionality grows exponentially with system size, and thus, complete exploration is out of the question. In addition, complex systems only sample parts of their space: they may get stuck in certain areas for very long periods of time, not exploring all available configurations. Important rearrangements, such as crossing between different regimes, occur but only infrequently and on timescales many orders of magnitude longer than typical dynamics. Developing reduced representations that capture the essential degrees of freedom while filtering out redundant details can alleviate both problems: tame the exponential growth of dimensionality and reveal collective coordinates along which transitions become accessible. However, the systematic, unsupervised determination of insightful representations or reduced representations remains an open problem. To overcome these difficulties, computational techniques and theoretical breakthroughs are required to represent and process spaces that are huge, structured, and only partially accessible (Coifman et al., 2005; Chodera et al., 2007).

To embody these concepts, this thesis turns to two systems that, in their radically different nature, encapsulate the spirit of complexity: proteins and learning networks. Both display an intricate interplay between structure and function, shaped

by the dynamics of many interacting components. For proteins as well as for artificial neural networks, large-scale behaviour arises from the exploration of a vast configurational space – one structured by molecular conformations, the other by patterns of weighted connectivity – offering complementary perspectives on how organisation and function emerge from collective dynamics. To address this, the present work combines methods from statistical physics, information theory, and computational modelling, aiming to connect the geometry of the configurational space to the emergent behaviour of these systems.

Proteins are natural macromolecules whose biological function arises (also) from their three-dimensional structure. They represent a canonical example of a system in which a vast configurational space underlies highly specific, reproducible behaviour. A polypeptide chain can, in principle, adopt an astronomical number of conformations, yet most naturally occurring proteins spontaneously fold into a stable native state within biologically relevant timescales. This apparent paradox – the emergence of a unique functional form through the exploration of an immense conformational space – has long been central to statistical and theoretical studies of folding (Levinthal, 1969; Dill et al., 1997). The first part of this thesis investigates how coarse but informative representations of the configurational space of proteins can capture the essential collective motions and pathways while reducing the complexity of microscopic description.

Learning networks, by contrast, are artificial systems designed to process information by adapting internal parameters. Despite their different nature, they share with biological systems key features such as high dimensionality, nonlinearity, and emergent collective behaviour. The training of a neural network can be viewed as a dynamical process exploring a configurational space defined by its weights or connectivity patterns. The geometry of this space – its minima, barriers, and flat regions – governs how the network learns, generalises, and responds to perturbations (Hochreiter et al., 1997a; Sagun et al., 2017). The second part of this thesis approaches these questions from a statistical physics perspective, aiming to characterise the organisation of the configurational space of learning networks and to relate its structural properties to their computational and generalisation capabilities.

This thesis is organised into two main parts, each devoted to one of the two classes of systems introduced above: proteins and learning networks. Following the general introduction, the first part focuses on proteins. It opens with a theoretical background chapter that introduces the main concepts of computational biophysics, *Molecular Dynamics*, *Coarse-Graining (CG)*, and the *Resolution-Relevance (RR)* framework, and is followed by three research chapters addressing, respectively, the relation between *RR* and informative reduced representations, the *CG* of protein conformations, and the identification of structurally optimal resolutions. The second part is devoted to learning networks. It begins with a theoretical chapter introducing feed-forward neural networks, the learning problem, *Wang-Landau* sam-

pling, and **Mapping Entropy (ME)**. It is followed by three research chapters on the multi-body decomposition of **WL**, the configurational space of neural networks, and the dropout space of reduced architectures. The manuscript closes with a general discussion of the main results and perspectives.

Proteins

2

Theoretical Background

Proteins are heterogeneous unbranched polymers made of a long sequence of building blocks, namely, the twenty-one proteinogenic amino acids (Ayon, 2020). All amino acids share the same structure; there is a central carbon atom (C_α) covalently bound to an amino group (NH_2), a carboxyl group ($COOH$), a hydrogen atom, and the R group, or side chain (Lopez et al., 2024). Consecutive amino acids are connected by a peptide bond formed between the carboxyl group of one residue and the amino group of the next, giving rise to a polypeptide chain with an amino (N) and a carboxyl (C) terminus. The side chain is the variable region of an amino acid; it differs widely in size and chemical properties, ranging from a single hydrogen atom in glycine to two aromatic rings in tryptophan, making it the largest and most hydrophobic residue. The properties of amino acids in solution are determined by the chemical characteristics of their side chain, as schematically reported in Figure 2.1.

The order of amino acids along a protein chain is not random: it is dictated by the genetic information encoded in the corresponding messenger RNA (mRNA) translated in the ribosomes. In this way, the genetic information converts into a linear polypeptide chain whose sequence carries all the instructions needed for the protein to reach its functional three-dimensional conformation (Morris et al., 2022). This principle is stated in the *central dogma of structural biology* (Crick, 1970): the amino acid sequence, also called the *primary structure* of the protein, encodes the spatial organisation and, consequently, its biological function.

The structure of a protein is organised in a well-defined hierarchy. Along the polypeptide chain, local interactions between backbone atoms give rise to the spontaneous formation of recurrent structural motifs. These local arrangements, stabilised mainly by hydrogen bonds, are collectively referred to as the *secondary structure* of the protein. Secondary structure elements can be broadly grouped into four categories: two regular and highly ordered motifs—the α -helix and the β -sheet—and two less structured ones, loops and turns, which connect the ordered regions and confer flexibility to the chain.

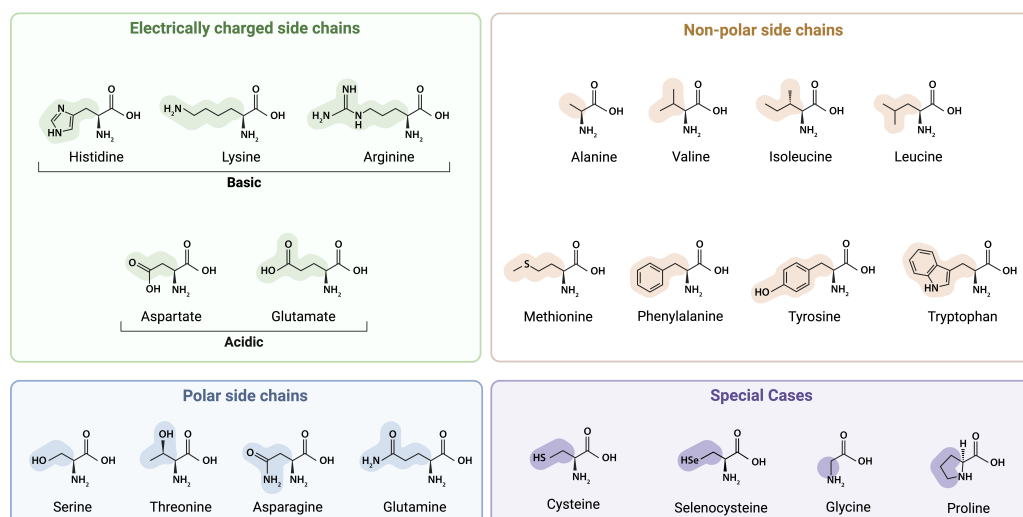


Figure 2.1: The twenty-one proteinogenic amino acids grouped according to side chain properties. Illustration created in *BioRender*.

The complete three-dimensional arrangement of these local elements defines the *tertiary structure* of the protein, which represents the stable fold adopted by the molecule in physiological conditions. Finally, in many cases, several folded chains associate through non-covalent interactions to form a higher-order assembly, known as the *quaternary structure*. This hierarchical organisation, from linear sequence to association of multiple subunits, is schematically illustrated in *Figure 2.2*.

Proteins play a central role in all biological processes, acting as enzymes, receptors, structural scaffolds, and molecular motors. Their function is closely linked to their structure: the specific three-dimensional arrangement of the polypeptide chain determines the precise task each protein performs within the cell. However, proteins are not static entities. Even in their native state, they fluctuate among closely related conformational ensembles, and these internal motions are often essential for their biological activity. Therefore, understanding this interplay between structure and dynamics is crucial for a complete description of protein function. In the next section, we will introduce the *Molecular Dynamics (MD)* approach, a pow-

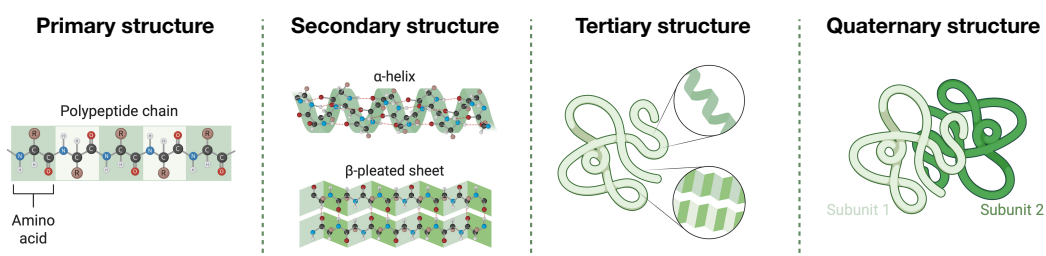


Figure 2.2: The four structural levels of a protein, from left to right: the sequence of amino acids (*primary structure*), the local folding into secondary structure elements (e.g., α -helices and β -sheets), the overall three-dimensional conformation of a single polypeptide chain (*tertiary structure*), and the assembly of multiple subunits into a complex (*quaternary structure*). Illustration created in *BioRender*.

erful computational framework for investigating protein conformational dynamics at the atomic level.

2.1 Molecular Dynamics

2.1.1 From a Quantum to a Classical Description

In the previous section, we introduced proteins from a biological perspective. However, from a physicist's point of view, a protein is nothing more than an assembly of atoms — mainly carbon, hydrogen, nitrogen, oxygen and sulphur — immersed in an aqueous and ionic environment (solvent) (Frenkel et al., 2023). A natural starting point for a theoretical description of such a system would be non-relativistic quantum mechanics, where the dynamics of both nuclei and electrons are governed by the many-body Schrödinger equation. In principle, this approach could be treated almost exactly using methods such as Density Functional Theory (E. Engel, 2011) or Quantum Monte Carlo (Nightingale et al., 1998). In practice, however, the enormous size of biomolecules and the long time scales relevant to their conformational dynamics make a full quantum-mechanical treatment computationally impractical. Moreover, weak and strong nuclear forces can be safely neglected as they become relevant only at energy and spatial scales far beyond those involved in biomolecular processes. For these reasons, atoms are typically modelled as classical point-like particles, and only the effective interactions arising from the underlying electronic structure are retained.

This simplification is formally justified by the Born-Oppenheimer approximation (Marx et al., 2009), which takes advantage of the large difference in mass between nuclei and electrons. Because electrons are much lighter, they respond almost instantaneously to any change in nuclear configuration. Consequently, the electronic energy can be treated as a function of fixed nuclear coordinates, effectively decoupling the fast electronic motion from the slow dynamics of the nuclei. Under this assumption, the electronic degrees of freedom are not explicitly considered, but rather implicitly integrated in the model, and the atomic nuclei are viewed as point-like particles whose positions and momenta evolve in time using the laws of classical mechanics. As a direct consequence, electronic excitations are neglected, and the bonding topology of the system is assumed to be fixed, meaning that chemical reactions and bond rearrangements are not allowed within this approximation.

2.1.2 Empirical Force Fields

The precise functional form of the interaction potential, or the associated *force field*, that governs the interaction between nuclei cannot be derived exactly from first principles (Ponder et al., 2003). Instead, it is postulated as an *ansatz* designed to capture the relevant physical interactions within a computationally tractable framework (Weiner et al., 1984). Each contribution is expressed in a predefined mathe-

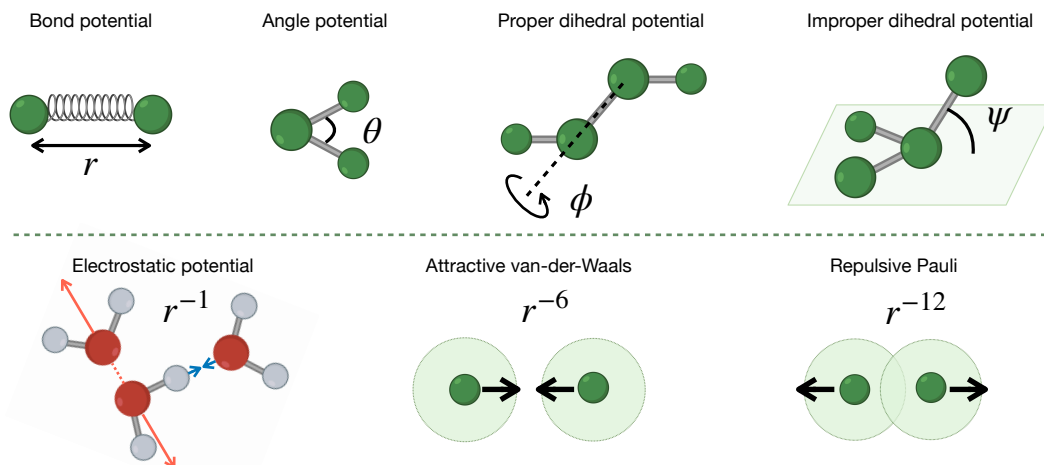


Figure 2.3: Schematic representation of the potential-energy terms commonly used in molecular-mechanics force fields. (**Top**) Bonded interactions: bond stretching (r), angle bending (θ), proper dihedral torsions (ϕ), and improper dihedrals (ψ). (**Bottom**) Non-bonded interactions: electrostatic (r^{-1}), attractive van der Waals (r^{-6}), and repulsive Pauli (r^{-12}) contributions. Illustration created in *BioRender*.

mathematical form with free parameters, which are subsequently tuned to reproduce experimental observables or high-level quantum-mechanical calculations. Over the past three decades, it has become standard practice to approximate these interactions using two major components (Cornell et al., 1995): a bonded potential and a non-bonded potential. The bonded part accounts for the interactions between atoms that are covalently linked; it is typically decomposed into bond stretching, angle bending, and dihedral torsion terms:

$$\begin{aligned}
 U_b = & \sum_{\text{bonds}} k_r (r - r_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \\
 & + \sum_{\text{dihedrals}} k_\phi [1 + \cos(n\phi - \delta)] + \sum_{\text{improper}} k_\psi (\psi - \psi_0)^2.
 \end{aligned} \tag{2.1}$$

The terms associated with bonds, angles, and improper dihedrals are usually described by harmonic potentials, reflecting the near-quadratic dependence of the energy on small deformations from the equilibrium geometry. Each term is characterised by an equilibrium value (r_0, θ_0, ψ_0) and a corresponding spring constant (k_r, k_θ, k_ψ), which quantifies the stiffness of the interaction. The proper dihedral term, instead, is typically expressed as a truncated Fourier series that captures the periodicity of torsional rotations and allows for multiple energetically favourable conformations. In most force fields, this series is limited to the first few terms (up to the fifth or sixth) to balance accuracy and computational cost.

The non-bonded potential describes the interactions between atoms that are not directly connected by covalent bonds. It accounts for the long-range electrostatic forces and the short-range van der Waals interactions, which together govern the packing, stability, and dynamical behaviour of molecular systems. In most

molecular-mechanics force fields, it is written as

$$U_{\text{nb}} = \sum_{ij} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{ij} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]. \quad (2.2)$$

The first term corresponds to the Coulomb interaction between partial charges q_i and q_j , separated by a distance r_{ij} and mediated by the vacuum permittivity ϵ_0 . The second term is a Lennard-Jones potential, which combines a steep repulsive component proportional to r_{ij}^{-12} , which represents Pauli exclusion at short distances, with an attractive dispersion term proportional to r_{ij}^{-6} . The parameters ϵ_{ij} and σ_{ij} define, respectively, the depth of the potential well and the finite distance at which the interatomic energy vanishes.

The interplay between these short-range repulsive and long-range attractive interactions defines the equilibrium structure of condensed matter systems, ensuring both molecular stability and the emergence of collective organisation. The different bonded and non-bonded terms of the force field are schematically illustrated in [Figure 2.3](#).

2.1.3 Integration of the Equations of Motion

Once an appropriate force field has been defined, the resulting potential energy function can be used to compute the forces acting on each atom, and consequently, to predict the time evolution of the entire system. This constitutes the basis of [MD](#) simulations, a computational technique in which classical equations of motion are numerically integrated to generate atomic trajectories. [MD](#) provides a bridge between microscopic interactions and macroscopic observables, enabling the investigation of structural, thermodynamic, and dynamical properties of molecular systems under controlled conditions. In general, [Molecular Dynamics](#) simulations can be formulated at different levels of description, ranging from fully quantum to mixed quantum–classical schemes. In the present context, however, we adopt the classical formulation introduced above, in which atomic nuclei are treated as point-like particles evolving on a fixed potential energy surface determined by the force field.

Within this framework, the time evolution of the system is governed by the Hamiltonian

$$H(\mathbf{r}, \mathbf{p}) = \sum_i^N \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}), \quad (2.3)$$

where $\mathbf{r}_i$ and $\mathbf{p}_i$ denote the position and momentum of the i -th atom, and $U(\mathbf{r})$ represents the total potential energy, given by the sum of the bonded and non-bonded contributions defined, respectively, in [Equation 2.1](#) and [Equation 2.2](#).

The Hamiltonian formalism provides a compact and general description of the

system's dynamics in the $6N$ -dimensional phase space defined by atomic coordinates and momenta. The equations of motion follow from Hamilton's canonical relations,

$$\begin{aligned}\dot{\mathbf{r}}_i &= \frac{\partial H}{\partial \mathbf{p}_i} \longrightarrow \dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i}, \\ \dot{\mathbf{p}}_i &= -\frac{\partial H}{\partial \mathbf{r}_i} \longrightarrow \dot{\mathbf{p}}_i = -\nabla_i U(\mathbf{r}).\end{aligned}\tag{2.4}$$

These relations express the time evolution of the atomic positions and momenta under the action of the interatomic forces derived from the potential energy surface. The second equation corresponds to Newton's second law, linking the acceleration of each atom to the negative gradient of the potential energy, $-\nabla_i U(\mathbf{r})$, which defines the force acting on particle i .

The complete set of $6N$ coupled differential equations describes the system's trajectory in phase space and, in principle, fully determines its temporal evolution once the initial positions and momenta are specified. In practice, however, an analytical solution is not accessible for systems of realistic size, as the potential energy surface depends on all atomic coordinates in a highly non-linear manner. The equations of motion must therefore be integrated numerically over discrete time steps, generating a sequence of configurations that approximates the continuous trajectory of the system. Several integration schemes have been developed for this purpose, among which the Verlet family—particularly the velocity-Verlet variant—has become standard for its simplicity, time reversibility, and excellent long-term energy conservation. These methods discretise time while preserving the symplectic nature of Hamiltonian dynamics, ensuring that the total energy remains bounded and stable over extended simulations.

The velocity-Verlet algorithm is obtained by truncating the Taylor expansion of the atomic positions and velocities at second order in time, yielding the update relations

$$\begin{aligned}\mathbf{r}_i(t + \Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2, \\ \mathbf{v}_i(t + \Delta t) &= \mathbf{v}_i(t) + \frac{1}{2}[\mathbf{a}_i(t) + \mathbf{a}_i(t + \Delta t)]\Delta t,\end{aligned}\tag{2.5}$$

where Δt is the integration time step, typically on the order of a few femtoseconds, and $\mathbf{a}_i(t)$ is the acceleration of atom i at time t , obtained from the interatomic forces divided by its mass.

2.1.4 Statistical Ensembles and Thermodynamic Control

In the absence of any external control, the numerical integration of Newton's equations conserves the total energy of the system, and the simulation samples the microcanonical (N, V, E) ensemble, where the number of particles N , the volume V and the total energy E are fixed. For biomolecular systems, however, this framework

is seldom appropriate: experiments are conducted at controlled temperature and often at ambient pressure; moreover, biomolecules are embedded in a fluctuating environment (solvent and ions) that exchanges energy and volume with the system and the surroundings. A consistent description, therefore, requires coupling the dynamics to external reservoirs to reproduce the canonical (N, V, T) or isothermal-isobaric (N, P, T) ensembles. In the canonical ensemble, the system can exchange energy with a thermal bath while keeping the number of particles and the volume fixed, so that the temperature remains constant on average. In the isothermal-isobaric ensemble, both energy and volume fluctuate as the system is additionally coupled to a pressure bath, maintaining constant temperature and pressure.

In the canonical ensemble (N, V, T) , configurations $\mathbf{x} = (\mathbf{r}, \mathbf{p})$ are distributed according to the Boltzmann factor:

$$P_{NVT}(\mathbf{r}, \mathbf{p}) = \frac{1}{Q_{NVT}} e^{-\beta H(\mathbf{r}, \mathbf{p})}, \quad Q_{NVT} = \int d\mathbf{r} d\mathbf{p} e^{-\beta H(\mathbf{r}, \mathbf{p})}, \quad \beta = (k_B T)^{-1}, \quad (2.6)$$

where Q_{NVT} is the canonical partition function, it acts as the normalisation constant of the distribution and encodes the thermodynamic properties of the system: all equilibrium quantities can, in principle, be derived from it. For example, the Helmholtz free energy is related to the partition function by

$$F = -k_B T \log(Q_{NVT}). \quad (2.7)$$

More generally, the same formalism allows one to compute *ensemble averages*, which represent the theoretical counterpart of experimentally measurable quantities at equilibrium. For any quantity $A(\mathbf{r}, \mathbf{p})$ the ensemble average is obtained as

$$\langle A \rangle_{NVT} = \frac{1}{Q_{NVT}} \int d\mathbf{r} d\mathbf{p} A(\mathbf{r}, \mathbf{p}) e^{-\beta H(\mathbf{r}, \mathbf{p})}. \quad (2.8)$$

In practice, the analytical evaluation of the partition function (Equation 2.6) or of ensemble averages (Equation 2.8) is intractable for systems of realistic size, even for relatively simple Hamiltonians. To estimate these quantities, one must rely on numerical approaches. In particular, *Molecular Dynamics* offers a practical route to compute ensemble averages without the explicit calculation of the partition function.

The connection is often described through the ergodic hypothesis, which states that long-time averages along a trajectory coincide with ensemble averages over the equilibrium phase-space distribution. This equivalence, however, is meaningful only when the dynamics explores the relevant region of phase space with the correct invariant measure and loses memory of its initial condition, that is, when it exhibits effective mixing. Under this hypothesis, time averages computed along a single MD trajectory are equivalent to ensemble averages over the canonical distribution, thereby linking the microscopic dynamics of the system to its macroscopic

equilibrium properties. Accordingly $\langle A \rangle_{NVT}$ can be estimated as:

$$\langle A \rangle_{NVT} = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(\mathbf{r}(t'), \mathbf{p}(t')) dt' \approx \frac{1}{M} \sum_{k=1}^M A(\mathbf{r}(t_k), \mathbf{p}(t_k)), \quad (2.9)$$

where $t_k = k\Delta T$ are the discrete time steps generated by the numerical integrator and M is the total number of sampled configurations. For sufficiently small sampling interval Δt , the discrete average provides an approximation to the corresponding continuous-time average. In the additional long-time limit $T = M\Delta t \rightarrow \infty$, this time average can in turn be identified with the ensemble average, provided that the MD simulation samples the canonical distribution, *i.e.*, under conditions of fixed N , V , and T . Although the number of particles and the volume can be easily controlled, maintaining a constant temperature during the simulation requires coupling the system to a thermal reservoir.

Researchers have developed several methods to couple the dynamics of the system to those of an external heat bath. These methods, collectively known as thermostats, are designed to regulate the system's temperature while preserving the statistical properties of the canonical ensemble. Although their specific formulations differ, all thermostat algorithms are integrated alongside the equations of motion and act to maintain a momentum distribution consistent with the *equipartition theorem*, which relates the average kinetic energy of the system to the temperature,

$$\langle K \rangle_T := \left\langle \sum_{i=1}^N \frac{\mathbf{p}_i^2}{2m_i} \right\rangle_T = \frac{3}{2} N k_B T, \quad (2.10)$$

where T is the target temperature, while the instantaneous value $T(t) = \frac{2K(t)}{3Nk_B}$ can fluctuate around it during the simulation.

Among the different approaches proposed, in this work we mainly adopted the velocity-rescale thermostat, a stochastic modification of the Berendsen scheme that combines the efficiency of velocity rescaling with accurate sampling of the canonical ensemble (Bussi et al., 2007). The method acts by rescaling all atomic velocities at each integration step through a factor λ that relaxes the instantaneous kinetic energy $K(t)$ toward its target value $K = \frac{3}{2} N k_B T$ over a characteristic coupling time τ_T . In contrast to the purely deterministic Berendsen approach, where the kinetic energy is forced to match its target exactly, the velocity-rescale scheme introduces a stochastic correction that allows thermal fluctuations consistent with the canonical distribution. As a result, the average temperature remains fixed, while the instantaneous temperature naturally oscillates around it with the correct statistical amplitude.

The coupling constant τ_T controls the strength of the interaction between the system and the thermal reservoir: large τ_T values correspond to weak coupling and nearly microcanonical dynamics, whereas small τ_T values produce faster thermali-

sation but stronger perturbations of the real dynamics. In this sense, the thermostat acts as a virtual heat bath that continuously exchanges small amounts of energy with the system, maintaining the desired temperature while preserving the canonical ensemble.

While the thermostat allows the system to exchange energy with an external heat bath and thus maintain a constant temperature, the volume has so far been treated as a fixed quantity. In realistic conditions, however—especially for solvated biomolecular systems—a system at constant pressure is subject to volume fluctuations. To reproduce these conditions in MD simulations and achieve a more faithful correspondence with experiments, one typically works in the isothermal-isobaric (N, P, T) ensemble, where the simulation box is allowed to fluctuate to sample the appropriate pressure-controlled equilibrium distribution. This ensemble represents a natural extension of the canonical formulation, in which the system is coupled not only to a thermal reservoir but also to a pressure reservoir through an additional control mechanism known as a barostat. The barostat regulates the system's pressure in analogy with how the thermostat controls its temperature, allowing the simulation to evolve at constant average temperature and pressure.

Formally, the probability of observing a configuration $\mathbf{x} = (\mathbf{r}, \mathbf{p})$ and an instantaneous volume V is given by

$$P_{NPT}(\mathbf{r}, \mathbf{p}, V) = \frac{1}{Q_{NPT}} e^{-\beta[H(\mathbf{r}, \mathbf{p}) + PV]}, \quad Q_{NPT} = \int dV d\mathbf{r} d\mathbf{p} e^{-\beta[H(\mathbf{r}, \mathbf{p}) + PV]}, \quad (2.11)$$

where P is the external pressure imposed by the reservoir and Q_{NPT} is the corresponding partition function.

At the microscopic level, the instantaneous pressure can be related to the interatomic forces through the work-virial theorem, which links macroscopic pressure to both kinetic and configurational contributions:

$$P(t) = \frac{Nk_B T(t)}{V(t)} + \frac{1}{3V(t)} \sum_i \mathbf{r}_i \cdot \mathbf{f}_i, \quad (2.12)$$

where $\mathbf{f}_i = -\nabla_i U(\mathbf{r})$ is the force acting on atom i .

Also in this case, as for the thermostat, several schemes have been developed to couple the system to an external pressure bath. Among them, we employed the Parrinello–Rahman barostat (Parrinello et al., 1981), a flexible-cell approach that extends the equations of motion by introducing the simulation cell vectors as additional dynamical variables. This formulation allows both the size and the shape of the simulation box to fluctuate in response to instantaneous internal stress, ensuring that the average internal pressure matches the target external value and that the system samples configurations consistent with the isothermal–isobaric (N, P, T) ensemble.

Together, these methodological components—accurate force fields, efficient integrators, thermostats, and barostats—form the foundation of contemporary MD

simulations (Frenkel et al., 2023; Karplus et al., 2002). They allow the controlled exploration of biologically relevant systems under thermodynamic conditions that closely mirror experimental setups (McCammon et al., 1988). Building on this robust foundation, MD has grown over the past several decades from a theoretical curiosity into a central method in computational biology. Since its birth in the late 1950s for atomic fluids (Alder et al., 1959) and its first application to biomolecules in the 1970s (McCammon et al., 1977), MD has become a fundamental tool for studying biomolecular systems. With continual advances in force field development, integration schemes, and hardware acceleration, MD now offers atomistic resolution on timescales ranging from nanoseconds to milliseconds (Hollingsworth et al., 2018). Software packages such as GROMACS (Abraham et al., 2015), coupled with massively parallel architectures (Stone et al., 2010), have made it possible to simulate complex systems like membrane proteins (Stansfeld et al., 2015), viral capsids (Freddolino et al., 2006), and large ribonucleoprotein assemblies (Villa et al., 2009). Specialised hardware platforms like Anton have further pushed the boundaries, enabling millisecond-long trajectories for biologically relevant systems (Shaw et al., 2008; Shaw et al., 2010).

Despite this progress, significant challenges remain, because many biologically relevant processes unfold on timescales—from microseconds to minutes and beyond—that sit far outside the temporal window that standard all-atom simulations can realistically capture (Karplus et al., 2002; Dror et al., 2012). These are rare events, typically involving substantial energetic or structural rearrangements that make them inherently difficult to observe within finite simulation times (Hänggi et al., 1990; Dellago et al., 2002; Bolhuis et al., 2002). As a consequence, the system may explore only a limited portion of its conformational landscape, potentially missing relevant transitions or yielding biased representations of the underlying dynamics (Pande et al., 2010; Tsabar et al., 2021).

To mitigate these intrinsic limitations, a wide range of methodological strategies have been developed, the most prominent ones being enhanced sampling techniques, Coarse-Graining, and multiscale modelling. Enhanced sampling aims to accelerate the exploration of configurational space by introducing carefully designed biases (Torrie et al., 1977), by exchanging replicas running under different thermodynamic conditions (Sugita et al., 1999), or by employing adaptive schemes that steer the system toward under-explored regions (Laio et al., 2002). Coarse-Graining (CG) and multiscale approaches address the complementary challenge of system size and complexity: by reducing the number of degrees of freedom or by combining multiple resolutions within a single simulation, they enable the study of larger assemblies and longer timescales than would be accessible with fully atomistic models (Izvekov et al., 2005; Warshel et al., 1976). In this work, we do not directly rely on these simulation techniques, and for this reason, they will not be introduced in detail here; comprehensive treatments can be found in standard reviews and textbooks (Voth, 2008).

All these methodological developments have undoubtedly expanded the reach of MD, constantly pushing the boundaries of the system size and timescales that can be explored. Yet MD remains a rapidly evolving field, and several open challenges persist. Among them, one stands out as a direct consequence of modern computational power: the sheer volume of data generated by high-resolution simulations (Husic et al., 2018; Noé et al., 2020). As simulations become longer and increasingly detailed, the resulting trajectories grow in dimensionality and complexity, often obscuring the underlying physical content. Therefore, extracting interpretable information from these datasets requires reducing the effective complexity—distilling essential collective variables while filtering out noisy or redundant degrees of freedom. This motivates a second, equally important role of CG: not as an *a priori* model reduction for simulation, but as an *a posteriori* analytical tool (Noid, 2013). In this perspective, CG provides a principled framework for identifying emergent structure in atomistic trajectories, uncovering relevant coordinates, and interpreting the dynamics at an appropriate level of abstraction. It is in this sense, and for this purpose, that CG will be introduced in the following section.

2.2 Coarse-Graining: From Multiscale Systems to Reduced Representations

2.2.1 Multiscale Structure and the Renormalization Group Perspective

In the previous sections, Molecular Dynamics was formulated at the atomistic level, where the state of a system composed of n particles is specified by a configuration $\mathbf{r} \in \mathcal{X} \subset \mathbb{R}^{3n}$, and evolves under an effective classical Hamiltonian. This representation is formally complete within the chosen level of theory. However, it is not necessarily the most informative description for the observables of interest.

Biomolecular systems are intrinsically multiscale: their macroscopic behaviour emerges from the collective organization of microscopic degrees of freedom, yet many experimentally accessible observables depend only on a restricted set of effective variables. This separation between microscopic detail and relevant large-scale behaviour motivates the search for reduced descriptions (L. D. Landau et al., 1980; Noid, 2013).

The conceptual framework that formalises this change of scale in statistical physics is the Renormalization Group (RG) (Fisher, 1998; Kadanoff, 1966; Wilson, 1975). In its original formulation, RG provides a systematic procedure to investigate how a physical theory transforms when progressively incorporating the effect of short-wavelength degrees of freedom. By integrating out microscopic fluctuations, one obtains an effective description governing the behaviour at larger length scales (Wilson, 1975; L. D. Landau et al., 1980). Near critical points, this procedure reveals scale invariance and universality, showing that macroscopic properties may become largely independent of microscopic details.

Although the systems considered in this thesis are not studied in the vicinity of criticality, the RG philosophy remains instructive. The central idea is that a detailed microscopic representation generally contains redundancy with respect to the observables under investigation. One can therefore construct an alternative representation in terms of effective degrees of freedom, designed to retain the relevant structural or thermodynamic features while reducing the dimensionality of the description (Noid, 2013; Shell, 2008).

In the context of molecular systems, this reduction corresponds to replacing the atomistic configuration by a set of collective variables, or CG sites. Formally, this operation acts as a scale transformation in configuration space: instead of tracking all microscopic coordinates, one introduces a reduced set of variables that encode the large-scale organisation of the system. Importantly, this is not merely a simplification of the interaction potential, but a transformation of the representational space itself.

From this perspective, CG is not only a computational strategy aimed at extending accessible time and length scales. It is a theoretical operation that reorganises the configuration space by grouping microscopic states into equivalence classes. In this sense, CG can be regarded as a representation-level analogue of the renormalization procedure: both involve the systematic elimination of degrees of freedom and the emergence of effective interactions at a larger scale.

In the following, we formalise this transformation and clarify its statistical-mechanical implications.

2.2.2 Formal Statistical–Mechanical Framework

In the canonical ensemble, as previously introduced, the equilibrium distribution in phase space is described by $P_{NVT}(\mathbf{r}, \mathbf{p})$ (Equation 2.6). However, the separability of the Hamiltonian, in kinetic and potential terms, implies that the kinetic contribution factorises and does not influence equilibrium structural observables. For this reason, the CG transformation can be formulated directly at the level of configuration space, in terms of the configurational marginal

$$p_r(\mathbf{r}) = \frac{1}{Z_X} e^{-\beta U(\mathbf{r})}, \quad Z_X = \int_X e^{-\beta U(\mathbf{r})} d\mathbf{r}. \quad (2.13)$$

The CG procedure consists in replacing the microscopic configuration $\mathbf{r} \in \mathcal{X} \subset \mathbb{R}^{3n}$ with a reduced set of variables describing the system at a lower resolution. Formally, this reduction is defined by a mapping

$$\mathbf{R} = M(\mathbf{r}), \quad (2.14)$$

where $\mathbf{R} \in \mathcal{S} \subset \mathbb{R}^{3N}$ with $N \leq n$. The mapping is a mathematical relation that associates the degrees of freedom of the high-resolution scale to the low-resolution ones, generally reducing the dimensionality of the representation.

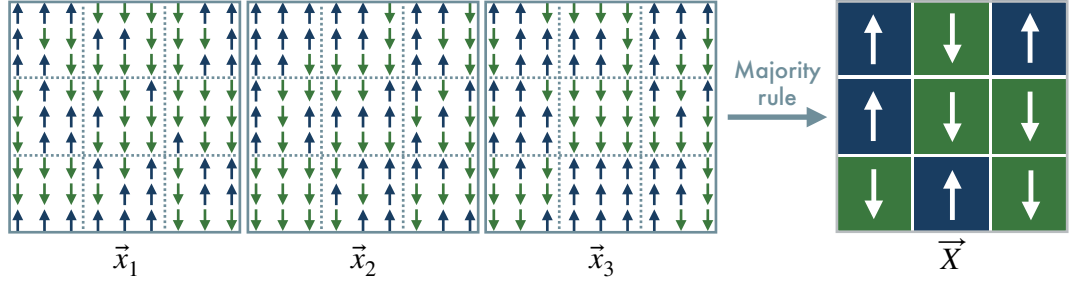


Figure 2.4: Schematic illustration of a block-spin coarse-graining transformation. Three distinct microscopic spin configurations $\vec{x}_1, \vec{x}_2, \vec{x}_3$ are mapped onto the same coarse configuration $\vec{X}$ through a majority rule applied within each block, $S_b = \text{sign}(\sum_{i \in b} s_i)$, with $s_i, S_i = \pm 1$. The transformation defines a many-to-one mapping in configuration space, highlighting the increase of degeneracy and the reduction of descriptive resolution.

A paradigmatic example is provided by the block-spin construction in lattice spin systems. In this case, groups of neighbouring spins $\{s_i\}$ are replaced by a single effective spin variable S_b , typically defined through a majority rule $S_b = \text{sign}(\sum_{i \in b} s_i)$. The coarse variable associated with each block represents the collective orientation of the underlying microscopic spins. This procedure, schematically illustrated in Figure 2.4, reduces the number of degrees of freedom while preserving large-scale features.

In general, the mapping $M : \mathcal{X} \rightarrow \mathcal{S}$ induces a partition of the microscopic configuration space into subsets

$$C_{\mathbf{R}} = \{\mathbf{r} \in \mathcal{X} \mid M(\mathbf{r}) = \mathbf{R}\}, \quad (2.15)$$

so that each coarse configuration $\mathbf{R}$ represents a family of microscopic states compatible with it. The mapping is therefore intrinsically many-to-one, and the dimensional reduction is accompanied by an increase of degeneracy at the coarse level. The many-to-one character of M implies that the mapping is not invertible: a given coarse configuration does not uniquely determine the microscopic configuration from which it originates. This loss of invertibility corresponds to a loss of microscopic information and is an intrinsic feature of any dimensional reduction. The integration over eliminated degrees of freedom therefore has structural consequences for the representation of the system.

The equilibrium probability of observing a coarse configuration $\mathbf{R}$ is obtained by summing the statistical weight of all microscopic configurations belonging to the corresponding subset $C_{\mathbf{R}}$. As schematically represented in Figure 2.5, the mapping M induces a redistribution of probability mass from the atomistic configuration space $\mathcal{X}$ to the reduced space $\mathcal{S}$. This corresponds to the marginalisation of the microscopic distribution over the degrees of freedom eliminated by the mapping

$$p_r(\mathbf{R}) = \int_{\mathcal{X}} p_r(\mathbf{r}) \delta(M(\mathbf{r}) - \mathbf{R}) d\mathbf{r}. \quad (2.16)$$

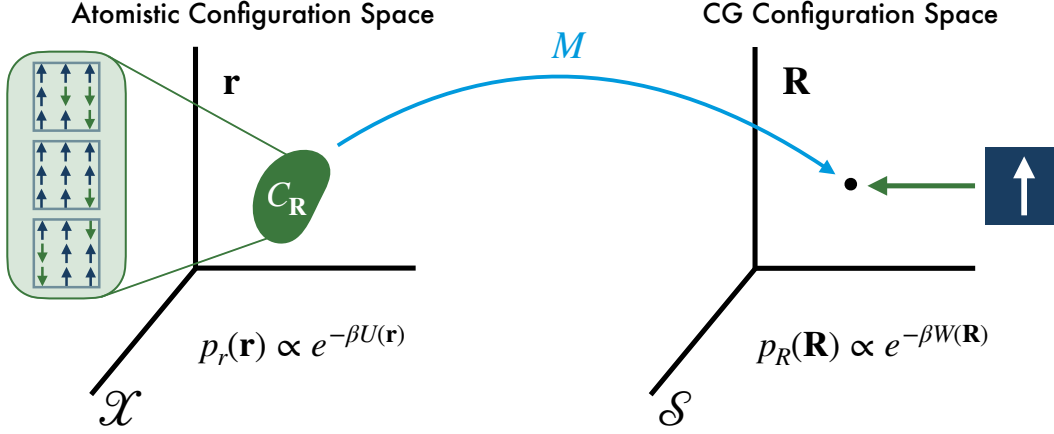


Figure 2.5: Schematic representation of the *Coarse-Graining* transformation between configuration spaces. The microscopic probability density $p_r(\mathbf{r}) \propto e^{-\beta U(\mathbf{r})}$ defined on $\mathcal{X}$ induces a reduced distribution $p_R(\mathbf{R}) \propto e^{-\beta W(\mathbf{R})}$ on $\mathcal{S}$ through the mapping $\mathbf{R} = M(\mathbf{r})$. The subset C_R represents the equivalence class of microscopic configurations $\mathbf{r}$ mapped onto the same coarse configuration $\mathbf{R}$, whose integration generates the many-body potential of mean force $W(\mathbf{R})$.

However, the *CG* probability distribution can be also expressed in canonical form by introducing an effective potential $W(\mathbf{R})$

$$p_R(\mathbf{R}) = \frac{1}{Z_S} e^{-\beta W(\mathbf{R})}, \quad Z_S = \int_{\mathcal{S}} e^{-\beta W(\mathbf{R})} d\mathbf{R}. \quad (2.17)$$

Imposing consistency between equilibrium probability distributions, *i.e.*, $p_r(\mathbf{R}) = p_R(\mathbf{R})$, leads

$$W(\mathbf{R}) = -k_B T \ln \left[\int_{\mathcal{X}} e^{-\beta U(\mathbf{r})} \delta(M(\mathbf{r}) - \mathbf{R}) d\mathbf{r} \right] + \text{const.} \quad (2.18)$$

The function $W(\mathbf{R})$ is the *Many-body Potential of Mean Force (MB-PMF)*. Despite its name, *MB-PMF* is not a mechanical interaction potential, rather a configuration-dependent free energy (Noid, 2013). It is termed “potential of mean force” because its gradient with respect to the coarse variables $\mathbf{R}$ yields the effective force acting on the reduced degrees of freedom. This force coincides with the conditional thermal average of the sum of the microscopic forces compatible with the imposed coarse configuration (Noid et al., 2008).

By construction, $W(\mathbf{R})$ depends parametrically on the thermodynamic state of the underlying high-resolution system. Since it is defined through the canonical measure, it inherits the dependence on temperature, volume, and particle number. Even if employed exactly, a *CG* model based on the *MB-PMF* is therefore strictly valid only at the state point at which it has been determined. The extension of a *CG* potential beyond its parametrisation conditions, and the construction of models capable of operating across different thermodynamic states, define the problem of *model transferability* (Noid, 2013; Noid, 2023).

At a fixed state point, the *MB-PMF* is in principle sufficient to ensure that the re-

duced model samples configuration space consistently with the underlying atomistic system. Two fundamental difficulties nevertheless arise. First, its evaluation requires the computation of a constrained partition function, which entails accurate sampling of microscopic configurations compatible with a given coarse state. Second, $W(\mathbf{R})$ is generally a genuine many-body function, depending simultaneously on all coarse coordinates. Even if the microscopic interaction $U(\mathbf{r})$ is pairwise additive, the integration over eliminated degrees of freedom generates effective many-body contributions. This structural complexity affects both its explicit representation and its practical use in simulations, and limits its straightforward applicability to systems differing in size or composition from those used in its derivation.

It is important to stress that the consistency discussed above concerns equilibrium configurational sampling. Even if a CG model reproduces exactly the canonical distribution of the underlying atomistic system, this does not imply dynamical equivalence. When expressed in terms of coarse variables, the dynamics of a deterministic microscopic system generally appears stochastic and non-conservative, due to the elimination of degrees of freedom. As a consequence, atomistic and Coarse-Graining models are typically characterised by different time scales, whose relation is non-trivial. In the following, the focus will remain on equilibrium structural properties and on the consistent sampling of configuration space.

2.2.3 Philosophies of Coarse-Graining

The formal construction developed in the previous section shows that the Many-body Potential of Mean Force $W(\mathbf{R})$ provides, in principle, an exact equilibrium description of the reduced system at a fixed thermodynamic state. In practice, however, the explicit evaluation of $W(\mathbf{R})$ is infeasible, and its inherent many-body character prevents its direct use in simulations, since computing it would require resolving high-dimensional, configuration-dependent interactions at every step. CG modelling therefore relies on approximate representations of the MB-PMF. The different *philosophies* of CG correspond to distinct strategies for constructing such approximations, depending on the type of information that is retained and on the criteria adopted to define consistency between microscopic and reduced descriptions. Historically, CG strategies are often grouped into two broad classes: *bottom-up* approaches, which derive effective interactions from a reference atomistic model, and *top-down* approaches, which parametrise interactions to reproduce macroscopic or experimental observables (Voth, 2008; Noid, 2013). More recently, a third class, *data-driven* and *knowledge-based* formulations, has become widespread, and many practical force fields adopt *hybrid strategies* combining multiple sources of information (Shell, 2008; Noid et al., 2024). In the following, we briefly recall representative methods from each class; the interested reader is referred to the cited reviews and original works for methodological details and practical implementations.

Bottom-Up

Bottom-up strategies determine effective interactions by reference to an underlying atomistic model, typically via the statistical analysis of MD trajectories or, more generally, equilibrium conformational ensembles. Their guiding principle is that the structural and thermodynamic properties of the reduced description should emerge systematically from the integration of microscopic degrees of freedom. Within this philosophy, one can distinguish three widely used methodological families: inverse approaches based on structural matching, force-based schemes, and variational formulations.

In practice, these families are often exemplified by (iterative) Boltzmann inversion, force matching, and relative entropy minimisation. (Iterative) Boltzmann inversion and related inverse techniques tune the parameters of a CG potential so as to reproduce selected structural observables of the atomistic reference, most commonly pair correlations such as radial distribution functions (Reith et al., 2003; Noid, 2013). Force matching instead determines the interaction parameters by minimising the discrepancy between the CG forces and the atomistic forces projected onto the chosen coarse variables (Izvekov et al., 2005; Noid, 2013). Finally, relative entropy minimisation provides a variational framework in which the CG parameters are obtained by minimising the information loss between the atomistic configurational distribution and the one induced by the reduced model within a prescribed functional family for the interactions (Shell, 2008; Noid, 2013). A more detailed treatment of relative entropy is provided in Chapter 6, where it is introduced in the context of the Mapping Entropy.

Top-Down

Top-down strategies parametrise CG interactions by targeting macroscopic, experimentally accessible observables, rather than by enforcing consistency with a specific atomistic reference model. Within this framework, the effective interactions are calibrated so that the reduced description reproduces selected thermodynamic and/or structural properties at the chosen resolution. Typical targets include bulk densities and compressibilities, equations of state, surface tensions, elastic moduli, and free energies of transfer between solvents. The functional form of the interaction is usually fixed *a priori* within a tractable Ansatz (e.g. predominantly pairwise additive potentials with limited angular dependence), and the parameters are then adjusted to match the selected observables. The calibration may be performed directly against experimental measurements, or against phenomenological constraints derived from continuum and statistical-mechanical considerations.

A canonical example in biomolecular modelling is the MARTINI force field (Marrink et al., 2007; Monticelli et al., 2008). In MARTINI, the non-bonded interaction strengths between CG beads are tuned to reproduce experimental free energies of transfer of small chemical fragments between aqueous and apolar envi-

ronments. These quantities, often inferred from water–octanol partition data, provide a thermodynamic proxy for hydrophobicity. Bonded interactions are subsequently chosen to preserve the local geometry at the adopted level of resolution. This construction yields a transferable and computationally efficient model that can be applied across a broad range of biomolecular systems. At a different level of resolution, structure-based models, such as Gō-like models (Takada, 1999; Taketomi et al., 1975) and elastic network models (Tirion, 1996; Bahar et al., 1997), constitute another established class of top-down constructions. In these approaches, the interaction network is defined directly from a known native structure, and the effective potentials are designed to stabilise that configuration while describing fluctuations around it. The objective is thus not to reproduce a full microscopic free-energy landscape, but rather to capture large-scale conformational motions consistent with experimentally resolved structures.

Top-down parametrisations are often designed for robustness across variations in temperature, concentration, or composition within a prescribed calibration domain. This is advantageous when atomistic reference data are scarce, or when predictive performance across different thermodynamic conditions is required. At the same time, because the parametrisation is constrained by a limited set of target observables, many other properties remain underdetermined. The resulting effective interactions may therefore entail compensating errors, and their connection to the exact many-body potential of mean force associated with a given mapping is, in general, indirect (Noid, 2013).

Knowledge-based

Knowledge-based strategies constitute a conceptually distinct class, although they share with top-down approaches the absence of an explicit atomistic reference trajectory for parametrisation. In this case, effective interactions are inferred from statistical analyses of structural datasets, most notably repositories of experimentally resolved biomolecular structures. The underlying assumption is that the observed frequencies of structural motifs encode effective preferences, which can be recast into *statistical potentials*. Typical examples include residue-residue contact potentials (Miyazawa et al., 1996), distance- and orientation-dependent statistical potentials (Sippl, 1990; Samudrala et al., 1998), and database-informed terms that bias secondary-structure preferences (Kabsch et al., 1983). The strengths of these models lie in their simplicity and in their ability to capture recurrent structural regularities; at the same time, their validity and transferability depend on the representativeness of the underlying datasets and on the approximations implicit in the statistical reconstruction (Russ et al., 2002).

Data-Driven

More recently, data-driven formulations have expanded the bottom-up paradigm by using machine-learning algorithms to approximate effective CG interactions within highly expressive functional classes. Neural-network-based CG force fields, including energy-based and equivariant architectures, can be trained to reproduce atomistic information through objectives closely related to force matching or variational distribution matching (J. Wang et al., 2019; Sahrman et al., 2025). Compared to classical analytical force fields, these models can capture complex many-body effects induced by CG; however, they introduce additional challenges in terms of data requirements, regularisation, interpretability, and extrapolation beyond the training conditions (Sahrman et al., 2025; Noid et al., 2024).

Hybrid strategies

In practice, many contemporary CG models adopt hybrid strategies that combine elements of bottom-up, top-down, and knowledge-based philosophies. For example, a model may enforce agreement with atomistic structural correlations while simultaneously matching experimental thermodynamic targets (e.g. pressure, density, surface tension), as commonly done by combining structure-based inversion schemes with thermodynamic corrections (Reith et al., 2003). Alternatively, one may incorporate database-derived structural biases within an otherwise bottom-up parametrisation, as in CG protein models that mix physically motivated terms with knowledge-based contributions (Davtyan et al., 2012). Such hybrid constructions reflect the need to balance microscopic consistency, empirical accuracy, and computational efficiency. From a formal standpoint, they can be interpreted as constrained approximations of the many-body potential of mean force associated with a chosen mapping, where different sources of information are weighted according to the intended scope of the model (Noid, 2013; Noid et al., 2024).

2.2.4 Resolution Level and Representational Power

The parametrisation strategy determines how the effective interactions are constructed, however, an equally fundamental modelling decision concerns the selection of the mapping itself. The CG map, $M : \mathcal{X} \rightarrow \mathcal{S}$, fixes the dimensionality of the reduced representation and, therefore, constrains the class of phenomena that can be faithfully described. The level of resolution is not merely a numerical simplification, but a structural transformation of the configuration space.

Recalling the definitions introduced above, the mapping M replaces $\mathbf{r} \in \mathcal{X} \subset \mathbb{R}^{3n}$ with $\mathbf{R} = M(\mathbf{r}) \in \mathcal{S} \subset \mathbb{R}^{3N}$, with $N \leq n$. The ratio N/n provides a *first* quantitative measure of dimensional reduction: decreasing N reduces the number of explicit degrees of freedom and generally enlarges the sets of microscopic configurations that are indistinguishable at the coarse level.

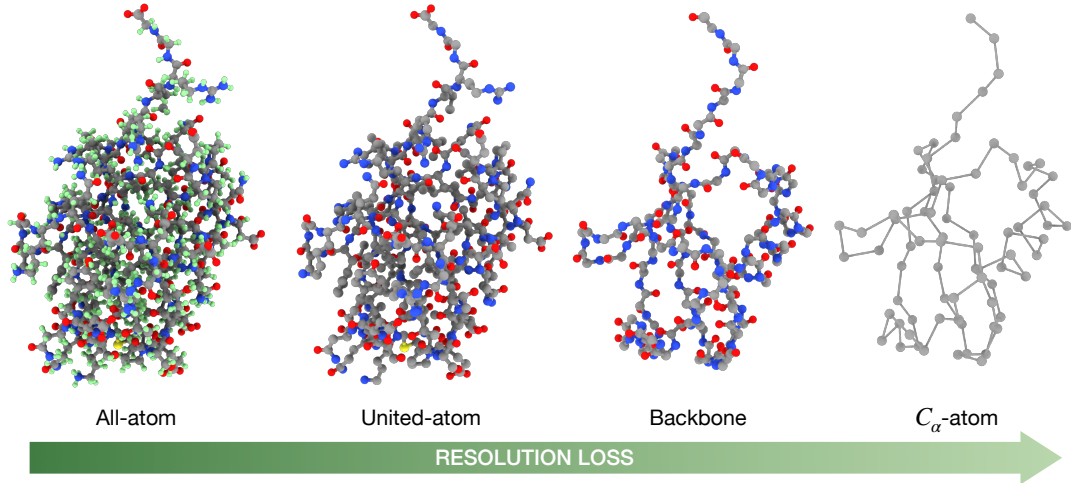


Figure 2.6: Hierarchical structural representations of the same protein at progressively decreasing resolution. From left to right: all-atom representation with explicit hydrogens, united-atom representation (hydrogens implicit), backbone-only representation, and C_α model (one interaction site per residue). Each representation corresponds to a different mapping $M : \mathcal{X} \rightarrow \mathcal{S}$, inducing a distinct reduction of configurational degrees of freedom and an associated increase in the degeneracy of microscopic states compatible with a given coarse configuration.

In biomolecular modelling, common choices range from united-atom representations, in which hydrogen degrees of freedom are integrated into heavy atom, to residue-level CG models where each amino acid is represented by one or a few interaction sites, up to ultra-CG descriptions where entire domains are treated as rigid bodies (see Figure 2.6 for a partial qualitative example). These alternatives correspond to mappings of progressively decreasing dimensionality and increasing configurational compression. The reduction in N is accompanied by an enlargement of the corresponding equivalence classes C_R (defined in Equation 2.15), since a growing number of microscopic configurations become indistinguishable within the reduced representation.

It is important to emphasise, however, that dimensionality alone does not adequately capture the notion of resolution. While decreasing N provides a quantitative measure of compression, the representational content of a model also depends on how microscopic degrees of freedom are grouped. Two mappings with identical N may preserve distinct structural features and thus differ in their descriptive power. The choice of mapping is therefore not reducible to selecting the number of retained variables, but requires identifying which aspects of the microscopic configuration space are relevant for the phenomena under investigation.

2.2.5 The Mapping Problem

Fixing the number of retained degrees of freedom sets a resolution level, but it does not uniquely determine the reduced description. At fixed dimensionality, different admissible mappings preserve different correlations and collective modes, and can

therefore lead to coarse models with distinct scope and performance.

It is useful to distinguish two modelling choices. The first is the *resolution choice*, namely how many effective variables are retained. The second is the *mapping rule*, namely how microscopic coordinates are grouped into effective sites and how coarse coordinates are defined. Even at fixed resolution, alternative rules can substantially modify the geometric embedding of configurations. For instance, centre-based projections, such as centres of mass and decimation schemes that retain selected atoms differ in which microscopic distinctions remain explicit and which are instead absorbed into the effective interactions.

A simple example at fixed resolution is a residue-level representation of a protein with one site per amino acid. Placing sites at the C_α positions preserves backbone geometry and fold topology directly, whereas defining each site as the residue centre of mass shifts the embedding and redistributes local fluctuations between explicit coordinates and effective interactions. Both choices retain the same number of variables, yet they encode different information and will generally privilege different classes of observables.

In practice, mapping rules are commonly selected through a combination of approaches. A first, widely used strategy is *heuristic* and chemistry-driven: atoms are grouped into beads representing chemically meaningful units, often with transferable bead types. This logic underlies, for example, MARTINI for lipids (Marrink et al., 2007; Marrink et al., 2013) and oxDNA for nucleic acids (Sengar et al., 2021; Poppleton et al., 2023). However, alternative choices for the same system can be equally well motivated, and physical considerations may compete with chemical intuition. A representative case is UNRES, where coarse centres are placed at peptide bonds rather than at the C_α atoms. This choice provides a convenient backbone parametrisation and shifts part of the local conformational detail into the effective interactions (Liwo et al., 1999; Liwo et al., 2005). A second, more physics-informed guideline is to choose groupings that align with the dominant organisation of interactions and motions, with the practical aim of obtaining effective interactions that admit simpler and more robust functional forms. If strongly coupled microscopic degrees of freedom are split across different sites, the reduced model must encode their correlations through increasingly intricate interactions, often many-body, which complicates parametrisation and reduces transferability (Rühle et al., 2009). A third, increasingly explored route is to treat the mapping itself as a *model parameter* to be selected by optimisation. In this perspective, each candidate mapping induces a different CG description and therefore a different amount of lost microscopic information. One can then endow the set of admissible mappings with an explicit objective functional, typically information-theoretic, that quantifies how well the reduced variables preserve the statistics of the atomistic ensemble. By evaluating this functional across alternative mappings, it becomes possible to rank alternative representations at fixed resolution and to identify those that retain more of the informative structure of the reference system, in the sense relevant to the

targeted observables, rather than relying solely on chemical intuition or geometric convenience (Foley et al., 2015; Diggins IV et al., 2019; Foley et al., 2020; Giulini et al., 2020; Giulini et al., 2021; Giulini et al., 2024; Menichetti et al., 2021).

These considerations clarify that a mapping is not merely a numerical reduction, but a choice of which microscopic distinctions remain accessible at the reduced level. The same trade-off between compression, retained information, and the descriptive usefulness of the remaining variables is systematised in the Relevance-Resolution framework introduced in the following sections (Marsili et al., 2013).

2.2.6 Coarse-Graining as an Analytical Operator

In the previous sections, CG was introduced mainly in its *constructive* role: once a resolution level and a mapping rule are fixed, one defines a reduced description and an effective interaction on the coarse variables. Here we stress a complementary use, which becomes particularly natural after the discussion of mapping degeneracy and information loss: the mapping can be applied *to existing atomistic trajectories* as an *analysis operator*, i.e. as a controlled lens that reorganises the data and exposes structured behaviour at a chosen level of abstraction (Noid, 2013).

Formally, a mapping $\mathbf{R} = M(\mathbf{r})$ induces a projection from the microscopic configuration space to a reduced one, partitioning microstates into equivalence classes labelled by R . The novelty of the analytical viewpoint is that this many-to-one character is not treated primarily as a modelling compromise, but as a deliberate tool: by deciding which microscopic distinctions remain explicit in R and which are integrated out, one selects the level at which patterns and regularities are sought. This projection redistributes complexity. The reduced distribution is obtained by marginalisation over the eliminated degrees of freedom, and the effective free-energy landscape on R collects both energetic and entropic contributions from all microstates compatible with the same coarse state. From an interpretative standpoint, this means that CG acts as a filter: variability orthogonal to the retained variables is absorbed into degeneracy, while collective features aligned with R are better defined and easier to characterise. As a consequence, reduced representations can improve robustness and interpretability when the microscopic description contains fast, noisy, or weakly relevant fluctuations for the questions at hand (Noid, 2013).

A practical consequence is that one can *compare* alternative or progressively coarser representations directly on the same reference to assess which one yields a more stable and informative organisation of the data for the observables of interest. This includes, for instance, checking the persistence of identified states or collective coordinates under changes of representation, and evaluating whether the resulting reduced description supports simpler statistical models or more robust estimates. Recent approaches make this idea quantitative by linking mapping choice to information-based criteria computed from the reference ensemble, thereby turning representation selection into a principled part of the analysis workflow (Diggins

IV et al., 2019; Foley et al., 2020; Shell, 2008).

In summary, beyond enabling faster simulations, **CG** can be viewed as an operator for *interpreting* high-dimensional molecular ensembles: it compresses the description in a controlled manner, suppresses selected microscopic variability, and can thereby reveal collective structure that is obscured at full resolution (Noid, 2013). This viewpoint is also the conceptual thread underlying the rest of this thesis: by treating the choice of representation as a tunable information filter, one can make explicit the trade-off between compression and retained content, and turn it into a quantitative object of study. In the following chapters, this idea is developed into criteria to compare representations across resolutions and to select those that preserve the information most relevant to the observables and processes of interest, providing a principled bridge between **CG**, inference, and model design (Marsili et al., 2013).

2.3 The Resolution-Relevance Framework

Coarse-Graining provides a systematic procedure for dimensional reduction of a complex system by combining microscopic configurations into a more manageable set of effective variables. As motivated in the previous chapter, these types of transformations are necessary to discover interpretable structure, suppressing microscopic fluctuations, and exposing large-scale patterns. However, even after establishing why **CG** is desirable, a central question remains unsolved: *how should one assess whether a particular representation is informative?* In the information-theoretic perspective adopted in this chapter, the level of detail at which a representation distinguishes among its states is quantified by its *resolution*. The resolution, introduced formally in the next section, reflects the average cost of encoding the distinctions present in the representation; however, a high level of detail does not by itself imply that these distinctions correspond to meaningful structure rather than fluctuations arising from limited sampling. A complementary measure is thus required to assess how informative a representation truly is.

This ambiguity becomes especially acute in the high-dimensional and under-sampled settings typical of modern biological and physical data. In these regimes, the number of variables is usually comparable to or even larger than the number of samples available, making a reliable estimation of statistical dependencies extremely challenging. Empirical frequencies fluctuate widely, entropy estimates become biased, and standard inference techniques require strong regularisation or restrictive modelling assumptions (Mézard et al., 2009; Grünwald, 2007). Such assumptions may inadvertently impose structures not supported by the data, obscuring the underlying generative process. The fundamental difficulty is that, under severe undersampling, only a limited number of data features can be considered reliable.

The **Resolution-Relevance (RR)** framework addresses this difficulty by intro-

ducing an intrinsic task-independent measure of how much information a dataset or representation carries about its unknown generative process (Marsili et al., 2022). Its starting point is the observation that, in the undersampled regime, the only statistically reliable quantity is the empirical frequency of observed outcomes; any higher-order statistic is dominated by noise. Frequency variation is therefore the natural basis for distinguishing meaningful structure from fluctuations. Representations that produce strongly heterogeneous frequency spectra must encode features that correspond to genuine underlying distinctions. Conversely, representations whose frequencies fall within a narrow range cannot be trusted to reflect meaningful structure.

Formalising this idea requires supplementing the traditional notion of resolution with a complementary quantity that reflects how strongly the data themselves support certain states of a representation over others. We refer to this quantity as *relevance*. Intuitively, relevance describes how unevenly the observations are distributed across the possible states: a representation is informative when some states consistently appear more often than others, signalling the presence of robust and reproducible structure. Conversely, when all states occur with comparable frequency, the representation offers little guidance in distinguishing meaningful patterns from noise, regardless of its level of detail. This definition is intrinsic—no external task or target variable is required—and it avoids the biases associated with parametric models or task-driven notions, such as those underlying the Information Bottleneck or supervised learning objectives (Tishby et al., 1999).

By jointly analysing resolution and relevance, the RR framework offers a systematic way to evaluate CG representations and to navigate the trade-off between retaining meaningful structure and suppressing noise. It provides information-theoretic constraints on what can be inferred from scarce data and it establishes a unified perspective on efficient representations in both physical and learning systems. The following sections develop the theoretical foundations of this framework, starting from the theoretical definition of resolution and then introducing relevance and its implications for statistical inference.

2.3.1 Resolution: Complexity and Coding Cost

The resolution of a representation quantifies how finely it distinguishes among the possible states of a system, or equivalently, how much microscopic variability is retained after CG. Let the microscopic configurations be denoted by $x \in \mathcal{X}$, with $\mathcal{X}$ an arbitrary configuration space. A mapping $\pi : \mathcal{X} \rightarrow \mathcal{S}$ assigns each configuration to a finite represented state $s \in \mathcal{S}$. The properties of the system, the structure of $\mathcal{X}$, and the physical interpretation of the representation are intentionally left unspecified; what matters for present purposes is the induced distribution:

$$p(s) = \sum_{x : \pi(x)=s} p(x) , \quad (2.19)$$

which records how frequently each represented state is observed under the underlying generative process. This distribution forms the starting point for quantifying the informational complexity of a representation in terms of coding cost.

The link between representational complexity and coding cost follows from classical information theory. In this setting, each represented state s is associated with a finite binary string, or *codeword*, that specifies its identity. For a lossless and unambiguous encoding, these codewords must be assigned in a way that allows the sequence of states to be uniquely reconstructed and as short as possible. This principle, made precise in Shannon's formulation of optimal compression (Shannon, 1948), naturally leads to the quantity

$$E_s = -\log p(s), \quad (2.20)$$

which represents the minimal number of nats required to specify state s . The average cost of this optimal description is the Shannon entropy

$$H[s] = -\sum_{s \in S} p(s) \log p(s), \quad (2.21)$$

which therefore provides a natural measure of the complexity, and hence the resolution, of the representation. Representations with high entropy finely partition the configuration space: many states occur with comparable probability and require non-trivial coding cost. Representations with low entropy correspond to coarse partitions in which many microscopic configurations are merged, and only a small number of states appear frequently.

A concrete illustration of these ideas can be obtained by considering the temporal discretisation of a sparse event sequence, such as the spike times of a neuron (Cubero et al., 2020). Leaving aside the biological context, the data reduce to a sequence of event times $\{t_1, \dots, t_M\}$ within an observation window of duration T . Different choices of temporal resolution correspond to different ways of partitioning this window into represented states. Dividing the interval into bins of duration Δt induces a mapping in which each event is assigned to one represented state—namely, the temporal bin it falls into. Varying Δt therefore provides an intuitive way to visualise how the fineness of a partition controls the number of distinguishable states and, consequently, the entropy of the representation. For small Δt , the observation window is divided into many narrow bins; most bins contain at most one event, and the entropy approaches its maximal value, reflecting a highly resolved description. For larger Δt , several events fall into the same bin, the number of distinct represented states decreases, and the entropy correspondingly drops. In the extreme case $\Delta t = T$, the entire observation window is treated as a single bin, all events correspond to the same represented state, and the entropy vanishes. This behaviour is summarised in Figure 2.7, which illustrates how progressively enlarging Δt groups multiple events into the same temporal region, thereby reducing the

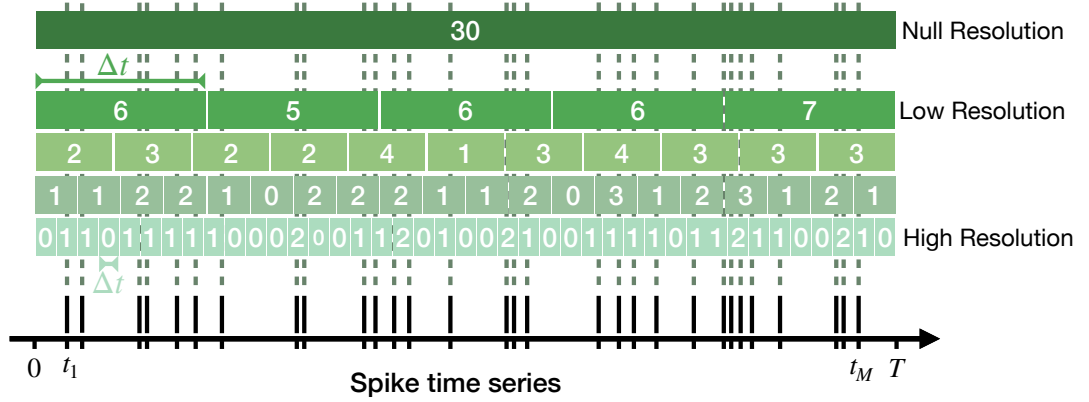


Figure 2.7: Illustration of how the resolution of a representation depends on the choice of the time bin Δt when discretising a spike time series. The same sequence of spike times (bottom) is grouped into temporal bins of increasing width. At high resolution (small Δt), the observation window is divided into many narrow bins, so that spikes are spread over a large number of distinct represented states and the entropy $H[s]$ is large. At low resolution (intermediate Δt), wider bins are used, and several spikes fall into the same temporal bin, reducing the number of represented states and lowering $H[s]$. At null resolution (maximal Δt), the entire observation window is treated as a single bin, all spikes correspond to the same represented state, and the entropy vanishes. The figure illustrates how progressively increasing Δt reduces the number of distinct temporal states and thereby lowers the resolution of the representation.

number of effectively distinguishable states.

The spike-time example shows that resolution is fully determined by how finely the mapping π partitions the underlying configuration space $\mathcal{X}$. Increasing the number of represented states raises the entropy $H[s]$ and yields a high-resolution description. This mechanism appears across modelling frameworks. In clustering, adding more clusters increases the granularity of the representation (MacQueen, 1967; Tibshirani et al., 2001); in latent-variable models, enlarging the latent space allows the system to encode more detailed configurations (Hinton et al., 2006; Kingma et al., 2014). Conversely, CG procedures deliberately *lower* resolution by collapsing many microstates into fewer macrostates. In that context, reducing resolution is not a defect but a modelling choice: it suppresses microscopic variability to highlight large-scale regularities that would otherwise be hidden (Kadanoff, 1966; Goldenfeld, 2018).

This observation leads to a crucial point. Resolution by itself does not guarantee that the distinctions introduced by a representation reflect meaningful structure in the data. Two representations may have the same entropy yet differ completely in the amount of information they reveal about the generative process. In high-dimensional, undersampled regimes, it is always possible to increase resolution by artificially subdividing the space into many poorly populated states (Marsili et al., 2013; Haimovici et al., 2015). Those refinements raise the entropy without necessarily improving, and often degrading, the interpretability and statistical robustness of the representation. A second quantity is therefore needed to assess whether a representation captures statistically supported structure in the data. This complementary

notion, *relevance*, is introduced in the next section.

2.3.2 Relevance: Informative Variability in Coding Cost

The resolution quantifies the average coding cost of a representation, but it does not specify which part of this cost reflects meaningful structure. Its value combines contributions that arise from systematic differences in how the data populate the represented states with contributions that are due solely to sampling fluctuations (Marsili et al., 2013; Marsili et al., 2022). This distinction is made explicit by the identity

$$H[s] = H[E] + H[s|E], \quad (2.22)$$

where E denotes the coding cost associated with state s . Since E is fully determined by s , the joint entropy can be written equivalently as $H[s, E] = H[s]$ and, leveraging the chaining rule, as $H[s, E] = H[E] + H[s|E]$, from which the decomposition follows. To understand their meaning, it is necessary to analyse how coding costs are distributed across states.

The probabilities $p(s)$ induce a distribution of coding costs,

$$p(E) = \sum_{s \in S} p(s) \delta(E - E_s) = W(E) e^{-E}, \quad (2.23)$$

where $E_s = -\log p(s)$, $\delta(\cdot)$ is a Kronecker delta, and $W(E)$ is the number of states with coding cost E . Treating both s and E as discrete variables for simplicity¹, the entropy of this distribution,

$$H[E] = - \sum_E p(E) \log p(E), \quad (2.24)$$

measures how broadly the coding costs vary across states. The second term in Equation 2.22 is

$$H[s|E] = \sum_E p(E) \log W(E), \quad (2.25)$$

which quantifies the degeneracy of the coding costs. Degeneracy arises when several states share the same coding cost; in such cases, the data do not provide evidence to distinguish among them, and the corresponding contribution is attributed to sampling noise. The decomposition, therefore, separates the informative component $H[E]$ from the noise component $H[s|E]$.

In this framework, the quantity $H[E]$ is identified as the *relevance* of the representation, since it captures the systematic, data-supported variation in coding costs that cannot be explained by sampling fluctuations alone.

¹ In the case where E is a continuous variable, the Equation 2.24 corresponds to the differential entropy of E (Cover et al., 2006). Because $p(E|s) = \delta(E - E_s)$ is a Dirac delta, the conditional differential entropy diverges to $-\infty$. This divergence can be avoided by introducing a finite resolution in the definition of entropy, as discussed in (Cover et al., 2006). A detailed treatment of the continuous case is provided in (Duranthon et al., 2021).

2.3.3 Resolution and Relevance in Practice

Together, resolution and relevance describe complementary aspects of a representation: resolution quantifies its overall complexity, *i.e.*, the average coding cost required to specify the represented states, while relevance isolates the portion of that complexity that is supported by the data through systematic variations in coding costs. This complementarity is made precise by the decomposition of the coding cost, but it becomes more evident when examined in a concrete example. In realistic scenarios, however, the underlying probability distribution from which the data are drawn is unknown, so the probabilities $p(s)$ that enter the definitions of resolution and relevance cannot be accessed directly. When only a finite dataset of N observations is available, the best approximation to $p(s)$ is the empirical distribution $\hat{p}(s) = k_s/N$, where the frequency $k_s = \sum_{i=1}^N \delta_{s(i),s}$ counts how many times the state s occurs in the sample. The empirical resolution is then

$$\hat{H}[s] = - \sum_s \hat{p}(s) \log \hat{p}(s) = - \sum_s \frac{k_s}{N} \log \frac{k_s}{N}, \quad (2.26)$$

which quantifies the coding cost associated with the specific sample under analysis. This quantity should not be interpreted as an estimate of the true entropy $H[s]$ of the unknown distribution $p(s)$, since this estimator is well known to be biased (Bonachela et al., 2008; Carlton, 1969), especially in the undersampled regime (Bonachela et al., 2008; Nemenman, 2011). Estimating $H[s]$ is not the objective here; the empirical entropy $\hat{H}[s]$ serves solely as a sample-based measure of resolution.

To characterise how the sample is distributed across frequencies, let m_k denote the number of states observed exactly k times in the dataset. The empirical distribution over frequencies is then

$$\hat{p}(K = k) = \frac{k m_k}{N}, \quad (2.27)$$

i.e., the probability that a randomly chosen observation belongs to a state with count k . The empirical relevance is defined as the entropy of this distribution,

$$\hat{H}[k] = - \sum_k \frac{k m_k}{N} \log \frac{k m_k}{N}. \quad (2.28)$$

To illustrate these ideas in a controlled setting, it is useful to consider an *in silico* system where the generative process is fully specified. Following Grigolon et al. (2016), one can construct datasets composed of binary strings

$$s = (s_1, \dots, s_L), \quad s_i \in \{0, 1\}, \quad (2.29)$$

drawn from a prescribed probability distribution designed to combine variables with qualitatively different statistical roles.

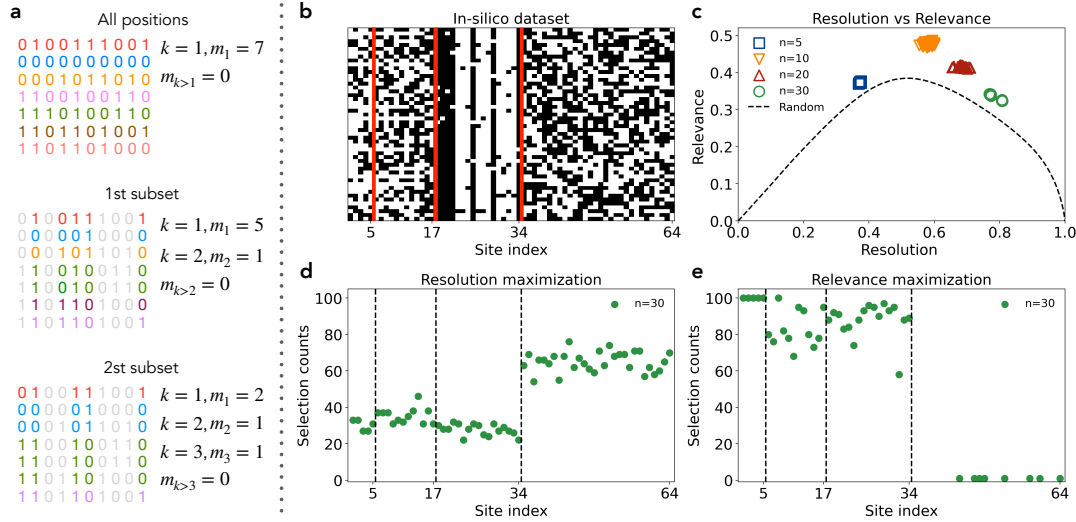


Figure 2.8: Illustration of relevance and resolution in an in-silico sequence dataset. (a) Example of three samples obtained from the same dataset by selecting different subsets of positions, illustrating the corresponding distinct states s . For each panel, the occurrence counts k and their multiplicities m_k are reported. (b) Example of a typical in-silico sequence dataset. The different regions composing each sequence (core, functionally dependent sites, conserved sites, and random sites) are highlighted as described in the main text. (c) Relevance $H[k]$ as a function of the resolution $H[s]$ for subsets of sites selected at fixed subsequence length n . The dashed curve corresponds to the random expectation obtained by placing n balls into L boxes while accounting for the number of occupied boxes. (d) Single-site selection counts $c_i(n)$ for subsequences of length $n = 30$ obtained by maximising the resolution $H[s]$, showing a preference for highly variable sites. (e) Single-site selection counts $c_i(n)$ for subsequences of length $n = 30$ obtained by maximising the relevance $H[k]$, highlighting the preferential selection of conserved and core sites.

The dataset is deliberately constructed to include heterogeneous classes of sites, not all of which are equally informative, in order to disentangle the effects of mere variability from those of structured statistical dependence. In particular, each sequence is composed of four distinct groups of variables, which are shown schematically in Figure 2.8.b and arranged from left to right according to their statistical role. The first group consists of a small core of sites generated so as to exhibit strong and non-trivial correlations, including dependencies that go beyond pairwise statistics. The second group is made of sites that depend hierarchically and noisily on the core variables, thereby propagating information in an indirect and distributed manner. The third group consists of highly conserved sites, which display little variability across the dataset and therefore contribute to resolution without carrying information about the underlying dependency structure. Finally, the fourth group is formed by independent random sites, which introduce variability but do not convey systematic information. A full technical specification of the data-generating process is reported in Grigolon et al. (2016).

By construction, not all sites contribute equally to the information carried by the sample. The analysis therefore proceeds by restricting attention to subsets of sites, and by treating the corresponding subsequences as the effective states of the system. Given a subset of positions $I \subseteq \{1, \dots, L\}$, each full configuration $s = (s_1, \dots, s_L)$ is

mapped onto a reduced state

$$s_I = \{s_i : i \in I\}, \quad (2.30)$$

so that configurations differing only on sites not included in I are mapped to the same reduced state. Formally, for a dataset of N sequences $\{s^{(\alpha)}\}_{\alpha=1}^N$, the empirical frequency of a reduced state s_I is

$$k_I(s_I) = \sum_{\alpha=1}^N \delta_{s_I, s_I^{(\alpha)}}, \quad (2.31)$$

that is, the number of times the subsequence s_I occurs in the sample. Different reduced states may occur with the same frequency; to characterise this degeneracy, we define the multiplicities

$$m_I(k) = \sum_{s_I} \delta_{k, k_I(s_I)}, \quad (2.32)$$

which count how many distinct reduced states are observed exactly k times.

The effect of this CG is illustrated in Figure 2.8.a. When all sites are retained, most configurations appear only once in the sample, resulting in a trivial frequency distribution concentrated at $k = 1$. Restricting the representation to a smaller subset of sites merges many distinct configurations into the same reduced state, generating a broader distribution of frequencies with non-trivial multiplicities. Importantly, different choices of the subset I —even when they contain the same number of sites—can induce qualitatively different frequency distributions, giving rise to distinct values of $\hat{H}[s_I]$ and $\hat{H}[k_I]$, as defined respectively in Equation 2.26 and Equation 2.27.

As illustrated in Figure 2.8.c, enlarging the subset of retained sites typically increases the resolution, $\hat{H}[s_I]$, because more degrees of freedom are available to distinguish sequences. However, the resolution is not fixed by the subset size alone: for a given number of retained sites $n = |I|$, different choices of I can lead to different values of $\hat{H}[s_I]$, depending on how variable the selected positions are in the sample. The relevance, $\hat{H}[k_I]$, is equally sensitive to which sites are retained: at fixed n it can vary widely across different subsets, and, as a function of n , it displays a non-monotonic behaviour. As the number of retained sites increases, the relevance initially grows, reaches a maximum, and eventually decreases when additional sites predominantly contribute variability without enhancing the heterogeneity of state frequencies.

This observation motivates comparing subsets of sites of fixed size n according to two distinct criteria. One may privilege subsets that maximise resolution, $\hat{H}[s_I]$, thereby favouring representations that distinguish as many reduced states as possible in the sample. Alternatively, one may privilege subsets that maximise relevance, $\hat{H}[k_I]$, favouring representations in which the observed reduced states

are distributed as heterogeneously as possible across frequencies. Although both criteria act on subsets of the same size, they emphasise different properties of the induced empirical distributions and, therefore, generally select different subsets of variables.

This distinction becomes explicit when analysing how frequently individual sites are selected across optimal subsets. To this end, we consider the single-site selection counts $c_i(n)$, defined as the number of times site i appears in an optimal subset of size n obtained by repeatedly applying a given optimization criterion. As shown in Figure 2.8.d, subsets that maximise resolution are dominated by sites with large $c_i(n)$ located in the final portion of the sequence, corresponding to the independent and identically distributed region of the dataset. These sites exhibit the largest fluctuations across samples and therefore contribute most effectively to increasing the number of distinguishable reduced states, largely irrespective of the underlying dependency structure.

By contrast, subsets that maximise relevance exhibit a markedly different selection pattern. As shown in Figure 2.8.e, sites belonging to the conserved region and to the correlated core are preferentially selected, while the random sites display negligible selection counts. These results demonstrate that the two criteria, despite operating on subsets of identical size, identify qualitatively different representations of the same dataset.

The *in silico* example thus makes it clear that resolution and relevance capture genuinely distinct aspects of a representation. While increasing resolution generally corresponds to resolving a larger number of distinct states, relevance quantifies how unevenly the available observations are distributed across those states. As a consequence, increasing the representational complexity does not automatically imply an increase in relevance, and the two quantities need not be optimised simultaneously.

2.3.4 The Resolution–Relevance Tradeoff

The previous section highlighted that increasing resolution and increasing relevance are, in general, competing objectives in the undersampling regime. Refining a representation introduces distinctions that are progressively less supported by data, while CG suppresses fluctuations but may also discard informative structure. The RR trade-off formalises this competition.

The interplay of these measures can be made explicit by recalling the decomposition of the coding cost introduced in Equation 2.22, where the resolution $H[s]$ is separated into an informative contribution, identified with the relevance $H[k]$, and a residual term $H[s|E]$ associated with degeneracy in the coding cost. This residual contribution reflects the fact that several represented states may occur with the same empirical frequency. Expressed in terms of the multiplicities m_k , this degeneracy

term takes the explicit form

$$H[s|E] = H[s|k] = \sum_k \frac{k m_k}{N} \log m_k, \quad (2.33)$$

which yields the following equivalent expression for the relevance

$$H[k] = H[s] - \sum_k \frac{k m_k}{N} \log m_k. \quad (2.34)$$

This relation explicitates how the information content of a representation is constrained by the distribution of empirical frequencies. Increasing the resolution $H[s]$ enlarges the number of distinguishable states, but it also tends to increase the degeneracy contribution $H[s|k]$ when many states are weakly populated. In this situation, additional descriptive detail does not necessarily correspond to a larger informative content, since part of the coding cost reflects sampling fluctuations rather than a genuine statistical structure.

For this reason, relevance is naturally considered as a function of the resolution across different **CG** levels of the same dataset. Empirically, $H[k]$ typically displays a non-monotonic dependence on $H[s]$: starting from coarse representations, increasing resolution initially enhances relevance because systematic heterogeneity in empirical frequencies becomes visible, whereas further refinements mainly generate sparsely populated states and reduce relevance. The resulting relevance–resolution curve thus provides a compact characterisation of how much statistically supported information can be extracted at different levels of description.

A useful benchmark for interpreting the relevance–resolution relation is obtained by considering a null model in which observations are assigned to represented states in the absence of any underlying statistical structure. In this situation, the empirical frequency spectrum arises solely from finite-sampling effects: distributing N observations randomly across a set of states inevitably produces heterogeneous occupancies, even when no genuine distinctions exist among the states. This heterogeneity generates a non-zero relevance, not because of intrinsic structure in the data, but as a consequence of purely combinatorial fluctuations associated with undersampling.

The relevance–resolution curve obtained under this random assignment, therefore, defines the baseline amount of relevance that can be expected in the absence of statistical organisation. An example of this random expectation is shown in **Figure 2.8.c**. Empirical curves that lie significantly above this reference indicate the presence of genuine structure in the data, whereas curves approaching it suggest that additional representational detail primarily captures sampling noise rather than informative distinctions.

This perspective naturally leads to the question of what is the maximal relevance compatible with a given resolution once finite-sampling constraints are taken into

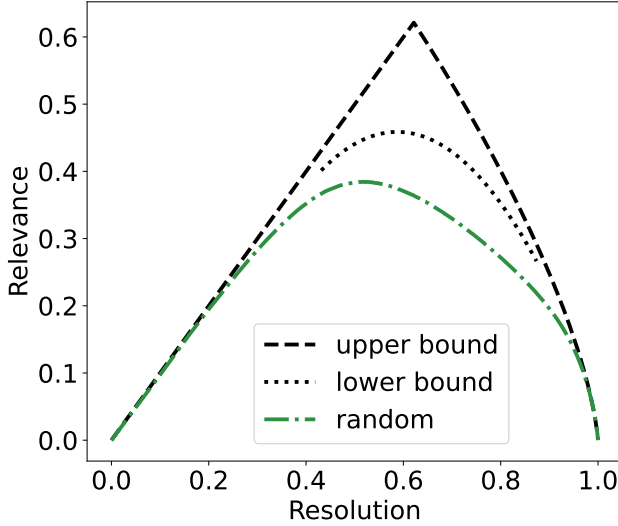


Figure 2.9: Relevance-resolution plane showing theoretical limits on the achievable relevance $H[\mathbf{k}]$ at fixed resolution $H[\mathbf{s}]$. The dashed black line denotes the upper bound to the theoretical maximum obtained from the continuous multiplicity approximation, the dotted black line the lower-bound to the theoretical maximum estimate based on stochastic multiplicities, and the green line the expectation corresponding to random assignments of observations to states.

account. Characterising this upper boundary, and the statistical signatures associated with it, is the focus of the next section.

2.3.5 Information-Theoretic Bounds

The considerations developed so far naturally lead to a further question. Once a given resolution is fixed, finite data constrain the admissible empirical frequency spectra; it is therefore legitimate to ask whether an upper bound exists on the relevance that can be achieved under such constraints. In other words, among all representations characterised by a given resolution $H[\mathbf{s}] = H_0$, which frequency spectra maximise the informative content captured by the relevance?

The admissible spectra are constrained by the normalization condition

$$\sum_{k \geq 1} k m_k = N, \quad (2.35)$$

which simply reflects the fixed sample size. Within this constraint, the resolution admits the explicit expression

$$H[\mathbf{s}] = - \sum_{k \geq 1} \frac{k m_k}{N} \log \frac{k}{N} = \log N - \sum_{k \geq 1} \frac{k m_k}{N} \log k. \quad (2.36)$$

At fixed N , prescribing the resolution therefore amounts to fixing the typical logarithmic frequency with which states are observed. Moreover, since the frequency k associated with a state s is a deterministic function of that state, the entropy decomposition introduced previously (see Equation 2.22) implies the general bound $H[\mathbf{k}] \leq H[\mathbf{s}]$. This reflects the fact that CG cannot increase the information content: retaining only the frequency class of a state necessarily discards part of the information contained in the original representation.

The determination of maximal relevance can thus be formulated as a constrained

variational problem over the multiplicities $\{m_k\}$. Introducing Lagrange multipliers, λ and μ , to enforce the normalization condition and the fixed-resolution constraint. The optimisation reduces to finding stationary points of the functional

$$\mathcal{F}[m_k] = H[k] + \mu \left(- \sum_{k \geq 1} \frac{k m_k}{N} \log \frac{k}{N} - H_0 \right) + \lambda \left(\sum_{k \geq 1} k m_k - N \right). \quad (2.37)$$

The solution to this problem is made non-trivial by the fact that the multiplicities m_k are non-negative integers. As a first step, it is nevertheless instructive to relax this constraint and treat m_k as continuous positive variables. This approximation provides an upper bound on the attainable relevance, since allowing fractional multiplicities enlarges the admissible space of frequency spectra beyond those compatible with a finite discrete sample. In equivalent terms, the optimisation can be formulated in terms of the intensive variables m_k/N ; for large datasets, the discreteness of the spectrum becomes negligible, as two adjacent admissible values differ by only $1/N$.

Within this relaxed setting, the optimisation reduces to a standard variational problem. Taking functional derivatives of $\mathcal{F}[m_k]$ with respect to m_k and imposing stationarity yields a family of frequency spectra whose asymptotic behaviour is of power-law form,

$$m_k \propto k^{-1-\mu}. \quad (2.38)$$

The emergence of a power-law multiplicity distribution has a clear information-theoretic interpretation. Maximising relevance at fixed resolution favours representations in which coding costs are broadly distributed across states. Uniform spectra would correspond to nearly degenerate coding costs and therefore carry little information about the underlying generative process. Conversely, sharply peaked spectra typically arise when the chosen resolution exceeds what finite data can reliably support, leading to many weakly populated states whose apparent distinctions are dominated by sampling noise. In contrast, maximally informative representations develop broad frequency spectra in which a few highly populated states coexist with a long tail of rarer ones, thereby maximising the entropy of the frequency classes compatible with the imposed resolution.

The fact that power-law multiplicity spectra are maximally informative can be illustrated through a simple analogy. Consider the task of locating a randomly chosen individual, within a population of N people. Without additional information, identifying that person requires $\log N$ bits. Suppose instead one knows the size k of the city where they lives, then one needs to identify one out of $k \cdot m_k$ individuals. The usefulness of this information depends on the distribution of city sizes. If all individuals live in a single city, knowing k provides no help. Similarly, if each city contains a single individual, the city size again carries no useful information, as

it is identical for everyone. Informative structure arises only when city sizes are broadly distributed, so that knowing k significantly restricts the search. Power-law distributions naturally realise this situation, with a few large cities coexisting with many smaller ones, thereby maximising the information conveyed by the frequency classes.

The exponent μ emerging from the variational solution admits a direct information-theoretic interpretation. Since it appears as the Lagrange multiplier enforcing the resolution constraint, it quantifies how the maximal achievable relevance varies when the coding cost is modified. More precisely, along the upper bound curve defined by the optimal frequency spectra one has

$$\mu = -\frac{\partial H[\mathbf{k}]}{\partial H[\mathbf{s}]}, \quad (2.39)$$

so that μ represents the (negative of the) slope of the relevance–resolution curve (see [Figure 2.9](#)). In information-theoretic terms, it measures how many bits of relevance are gained (or lost) for a marginal change of one bit in resolution. Varying μ therefore parametrises the family of optimal spectra that define the upper bound.

A particularly important regime arises when $\mu = 1$, where the gain in relevance exactly compensates the loss in resolution. At this point the upper bound curve reaches its maximal value compatible with the imposed constraints, identifying the regime of maximally informative representation. In this situation

$$m_k \sim k^{-2}. \quad (2.40)$$

This condition corresponds to Zipf’s law in the ranked frequency representation ([Zipf, 1932](#)). Indeed, a multiplicity spectrum $m_k \sim k^{-2}$ implies that the probability of observing a state with frequency k decays approximately inversely with its rank, a behaviour widely documented across biological, linguistic and learning systems. Within the present framework, the emergence of Zipf-like statistics is not a coincidence but rather a consequence of maximising the informative content of a representation under finite sampling constraints ([Schwab et al., 2014](#); [Aitchison et al., 2016](#)).

From a statistical-mechanical perspective, the emergence of Zipf’s law can be interpreted as a form of statistical criticality ([Schwab et al., 2014](#); [Mora et al., 2011](#)). Writing the coding cost as $E = -\log(k/N)$, a power-law multiplicity spectrum corresponds to an approximately exponential density of states in E . This exponential form reflects the underlying scale-free organisation of the frequency distribution, rather than the presence of a typical frequency scale. In the present framework, however, such critical-like statistics arise as a consequence of maximising informative content under finite sampling constraints rather than from fine tuning of external control parameters. A detailed and formal discussion of this connection can be found in [Marsili et al., 2013](#); [Haimovici et al., 2015](#); [Cubero et al., 2019](#).

As already discussed, relaxing the discrete constraint on the multiplicities m_k provides an *upper* bound on the maximal relevance, since treating m_k as continuous variables enlarges the admissible space of frequency spectra beyond those compatible with a finite dataset. A complementary strategy can be developed by keeping the multiplicities explicitly discrete and modelling them as random variables. This leads to a *lower* bound approximation to the theoretical maximum relevance, obtained by maximising the expected relevance over ensembles of admissible spectra rather than over individual deterministic profiles (shown as the dotted curve in Figure 2.9).

Concretely, the multiplicities are treated as stochastic counts fluctuating around a mean profile $n_k \equiv \mathbb{E}[m_k]$. A convenient analytically tractable assumption is to model these fluctuations as Poissonian,

$$m_k \sim \text{Poisson}(n_k), \quad (2.41)$$

which preserves discreteness while allowing controlled statistical variability. The optimisation then reduces essentially to evaluating the expectation $\mathbb{E}[m_k \log m_k]$. Its explicit derivation involves standard but rather technical steps that do not add conceptual insight here, and we therefore refer the reader to [Haimovici et al., 2015](#) for full details.

For our purposes, it is convenient to define

$$L(n) = \mathbb{E}_{m \sim \text{Poisson}(n)}[m \log m], \quad (2.42)$$

the stationarity condition determining the optimal spectrum takes the implicit form

$$L'(n_k) = A + \mu \log \frac{k}{N}, \quad (2.43)$$

where A collects constants associated with the normalisation constraint, while μ enforces the prescribed resolution constraint, as already discussed in the upper-bound derivation.

Despite the different optimisation strategy, the resulting spectra remain broad and approximately scale free. In particular, for large counts one has $L(n) \simeq n \log n$, which implies asymptotically

$$n_k \propto k^{-1-\mu}. \quad (2.44)$$

Thus, power-law multiplicity distributions arise robustly even when discreteness is explicitly enforced, showing that scale-free frequency spectra are not an artefact of the continuous relaxation but rather a generic feature of maximally-informative representations under finite sampling constraints.

Together, the upper and lower bounds on the theoretical maximum relevance delimit the admissible region in the $(H[s], H[k])$ plane that characterises the range

of informative representations compatible with finite sampling.

2.3.6 Multiscale Relevance

The analysis developed so far characterises the informative content of a representation at a fixed level of description. In many complex systems, however, statistical structure is distributed across multiple descriptive scales, and no single CG level is sufficient to capture all relevant regularities. This observation motivates extending the relevance–resolution framework to a multiscale setting (Marsili et al., 2013; Cubero et al., 2020).

Consider a configuration space $\mathcal{X}$ and a family of representations defined by mappings

$$\phi_\lambda : \mathcal{X} \rightarrow \mathcal{S}_\lambda, \quad (2.45)$$

parametrised by a scale parameter λ . Increasing λ corresponds to progressively coarser descriptions, obtained for example by aggregating variables, merging states, or reducing observational precision (Kadanoff, 1966). Each mapping induces a distribution $N_\lambda(M)$ over represented states $M \in \mathcal{S}_\lambda$, from which the corresponding resolution $H[s]_\lambda$ and relevance $H[k]_\lambda$ can be computed according to the definitions introduced previously.

As the CG level varies, the pair $(H[s]_\lambda, H[k]_\lambda)$ traces a curve in the RR plane. This curve summarises how the informative content accessible from the data changes with the descriptive scale. Fine-grained representations typically exhibit high resolution but low relevance due to sparse sampling, while excessively coarse descriptions suppress variability and again reduce relevance. Intermediate scales often reveal maximal heterogeneity in empirical frequencies, signalling statistically robust structure (Marsili et al., 2013; Cubero et al., 2020).

This observation motivates defining a global measure of informative content across scales. The *multiscale relevance* is naturally quantified by the area under the relevance–resolution curve,

$$\mathcal{R}_{\text{msr}} = \int H[k] dH[s], \quad (2.46)$$

where the integral is taken along the curve generated by the family of representations $\{\phi_\lambda\}$. This quantity captures the cumulative amount of statistically supported information that can be extracted from the data across all accessible levels of description. Representations whose high relevance value persists over a broad range of resolutions yield larger multiscale relevance, indicating robust hierarchical organisation, whereas datasets dominated by noise produce curves concentrated near the random baseline and therefore smaller areas. Conceptually, this construction is related to multiscale entropy approaches in complex systems, where informative structure is assessed across observational scales (Costa et al., 2005).

In practice, multiscale relevance can be estimated by constructing nested fami-

lies of representations—for instance by varying clustering granularity, discretisation resolution, or subsets of retained variables—and computing the empirical quantities $\hat{H}[s]_\lambda$ and $\hat{H}[k]_\lambda$ at each level. The resulting discrete **RR** curve provides a numerical approximation of $\mathcal{R}_{\text{msr}}$ through standard quadrature procedures.

Overall, multiscale relevance provides a synthetic characterisation of how informative structure is distributed across descriptive scales. Rather than focusing on a single optimal **CG**, it quantifies the robustness of statistical organisation under progressive changes of resolution, thereby offering a principled tool to assess hierarchical structure in complex datasets.

3

The Bliss of Dimensionality

3.1 Introduction

In the previous chapter, **Coarse-Graining (CG)** was introduced as both a practical strategy for reducing the dimensionality of complex molecular systems and an analytical operation for filtering microscopic variability and reorganising data at a chosen level of resolution (Shell, 2008; Noid, 2013). From this perspective, the central issue is not only *how to construct* a reduced representation, but *how to determine* whether that representation is genuinely informative of the underlying probability structure. The **Resolution-Relevance (RR)** framework provides a possible answer to this question by offering an intrinsic task-independent approach that quantifies the trade-off between descriptive detail and a statistically reliable structure directly from the sampled data, without the need for an explicit model of the underlying generative process (Marsili et al., 2013; Marsili et al., 2022).

This question is far more general than the molecular setting in which it is introduced in this thesis. In statistical physics, data analysis, and machine learning, it is often necessary to replace a continuous or high-dimensional description with a reduced representation that can be analysed, interpreted or sampled more effectively (Shannon, 1948; Bellman, 2015; Wainwright, 2019; Poma et al., 2025). Whenever a continuous or high-dimensional dataset is replaced by a discrete or reduced description, one must choose the level of resolution at which the representation is defined.

If the description is too coarse, relevant structure is lost, whereas if it is too fine, finite-sample fluctuations are amplified and statistically unreliable patterns emerge. This conflict is particularly acute in undersampled and high-dimensional regimes, where empirical frequencies provide only a partial and noisy view of the underlying

This chapter is based on material reported in M. Mele, D. Campos Moreno, and R. Potestio, *The bliss of dimensionality: how an unsupervised criterion identifies optimal low-resolution representations of high-dimensional datasets*, arXiv:2603.05214 (2026). At the time of writing, this work is publicly available as an arXiv preprint and has not yet been formally published. A public version is available at: <https://arxiv.org/abs/2603.05214>.

distribution (Bonachela et al., 2008; Nemenman et al., 2001).

In many applications, the level of resolution is fixed a priori, based on physical intuition, computational convenience, or established conventions. When a reference distribution is available, the level of detail can also be selected in a supervised way by maximising fidelity to that target, for example, through likelihood-based criteria or divergence minimisation (Kullback et al., 1951; Bishop, 2006). In many real problems, however, the distribution one would like to preserve is precisely what must be inferred from the data themselves (Awaya et al., 2024). This is the regime in which the RR framework is most appealing, but also the one in which its practical meaning must be established more carefully. (Haimovici et al., 2015; Duranthon et al., 2021; Marsili et al., 2022).

The aim of this chapter is therefore to test, in controlled supervised settings, whether the characteristic optima of the RR curve, namely the point of maximum relevance and the information-theoretic point at which the local slope reaches -1, identify, or at least robustly bracket, the representation that is optimal in terms of fidelity to a known generative probability distribution (Duranthon et al., 2021). To this end, RR is compared with a supervised benchmark based on the Kullback-Leibler (KL) divergence (Kullback et al., 1951). The analysis progresses from fully synthetic datasets, where dimensionality and latent structure can be tuned independently, to semi-real datasets constructed as Gaussian clones of MNIST (Deng, 2012), and finally to molecular dynamics trajectories of a simple benchmark molecular system, alanine dipeptide (Smith, 1999; Bolhuis et al., 2000). This progression makes it possible to assess the criterion first in settings with exact ground truth, then in increasingly structured datasets, and finally in a molecular context directly connected to the configurational CG problems discussed in the following chapters.

3.2 Materials and Methods

For each dataset considered in this chapter, we construct a family of discrete representations with varying numbers of states n . Each representation is then evaluated from two complementary viewpoints: its position on the RR plane, determined by the occupancies of the induced partition, and its fidelity to a known reference probability distribution, quantified through the KL divergence. The methodological comparison therefore consists in relating the two characteristic optima identified by the RR analysis to the representation that minimises the KL divergence. The present section introduces the construction of the representations, the quantities used for the comparison, and the datasets on which the analysis is performed.

3.2.1 Discrete Representations

A discrete representation is obtained by partitioning the sampled configurations into a finite number of states. Consider a dataset of P observations $\{x_i\}_{i=1}^P$, with

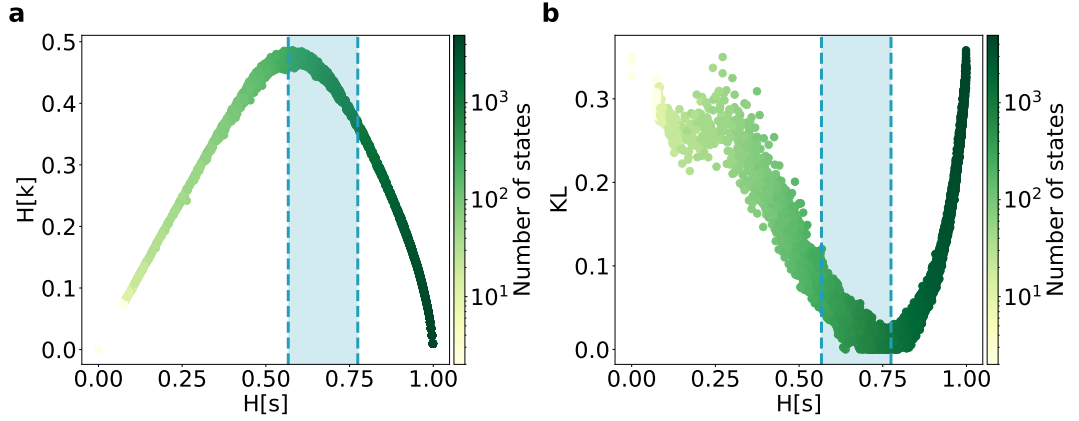


Figure 3.1: Schematic comparison between the unsupervised *RR* criterion and the supervised *KL*-based criterion. The figure shows the same family of discrete representations, parametrised by the number of states n (color scale), analysed from two complementary perspectives. (a) Representations in the *RR* plane. (b) Corresponding *KL* divergence as a function of the resolution for the same representations. In both panels, the shaded region marks the *RR* optimality region, delimited by the two characteristic estimates $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$.

each configuration x_i belonging to a space $\mathcal{X}$ of dimension N . A representation is defined by a mapping that assigns each observation to one of n states, thereby inducing a partition $\{C_j\}_{j=1}^n$ of the dataset, where C_j is the set of samples assigned to state j . The occupancy of state j is then $k_j = |C_j|$, with $\sum_{j=1}^n k_j = P$.

The procedure used to construct the representations depends on the dimensionality of the dataset. In low-dimensional settings, where the data space can be directly discretised, representations are built through axis-aligned N -dimensional histograms (Scott, 1979). The range of each coordinate is divided into a fixed number of intervals, generating a regular grid of hyper-rectangular cells. Observations are then mapped onto the grid, and the occupied cells define the states of the representation.

In higher-dimensional settings, where histogram-based discretisation becomes impractical because the number of bins grows exponentially with dimensionality (Bellman, 2015; Bishop, 2006), representations are instead constructed through agglomerative clustering (Murtagh et al., 2012; Freedman et al., 1981). Starting from singleton clusters, clusters are iteratively merged according to their pairwise distances until a prescribed number of clusters n is reached. The resulting clusters define the states of the representation, and their occupancies are given by the corresponding cluster sizes.

3.2.2 Relevance-Resolution and Supervised Optimality

For each discrete representation, the occupancies $\{k_j\}_{j=1}^n$ determine both the position of the representation in the *RR* plane and the empirical probability distribution associated with the partition. The same representation can therefore be evaluated from two complementary perspectives: an unsupervised one, based on the balance

between relevance and resolution, and a supervised one, based on the discrepancy between the empirical and the reference probability distributions. The comparison between the two criteria is schematically illustrated in [Figure 3.1](#).

Within the [RR](#) framework, each representation is characterised by the values of $(H_{\text{res}}$ and $H_{\text{rel}})$, computed from the empirical frequencies associated with the partition according to [Equation 2.26](#) and [Equation 2.27](#). By varying the representation size, one obtains a family of points in the $(H_{\text{res}}, H_{\text{rel}})$ plane, which defines the corresponding [RR](#) curve. As discussed in [Section 2.3](#), two characteristic representations are extracted from this curve. The first is the maximum-relevance point, denoted by $n_{\text{opt}}^{\text{MR}}$, corresponding to the representation at which $(H_{\text{rel}}$ is maximal. The second is the information-theoretic point, denoted by $n_{\text{opt}}^{\text{IT}}$, defined as the point at which the local slope of the [RR](#) curve reaches -1 ([Marsili et al., 2013](#); [Duranthon et al., 2021](#)). Since the curve is obtained from a discrete set of representations, the identification of these characteristic points requires smoothing and interpolation; the corresponding operational details are reported in [Appendix A](#).

To compare these unsupervised estimates with a supervised benchmark, each representation is also associated with an empirical probability distribution induced by the partition. Given a partition $\{C_j\}_{j=1}^n$ with occupancies $k_j = |C_j|$, the empirical probability of state j is

$$p_j = \frac{k_j}{P}. \quad (3.1)$$

The corresponding state probability is then assigned to each sample $x_i \in C_j$

$$\tilde{p}(x_i) = p_j = \frac{k_j}{P}. \quad (3.2)$$

Because the same state weight is assigned to all samples belonging to the same state, the resulting weights are not normalised over the dataset. A pointwise normalisation is therefore imposed, yielding the empirical distribution

$$\hat{p}(x_i) = \frac{\tilde{p}(x_i)}{\sum_{\ell=1}^P \tilde{p}(x_\ell)} = \frac{k_j}{\sum_{r=1}^n k_r^2}, \quad x_i \in C_j, \quad (3.3)$$

which satisfies $\sum_{i=1}^P \hat{p}(x_i) = 1$.

The empirical distribution $\hat{p}(x_i)$ is compared with a reference probability distribution $p_{\text{ref}}(x)$, assumed to be known or otherwise independently defined for the system under consideration. To obtain a discrete reference distribution on the sampled configurations, the reference density is evaluated at each observed point and then normalized over the dataset

$$p(x_i) = \frac{p_{\text{ref}}(x_i)}{\sum_{\ell=1}^P p_{\text{ref}}(x_\ell)}. \quad (3.4)$$

The discrepancy between the empirical and the reference distributions is quan-

tified through the **Kullback-Leibler** (KL) divergence,

$$D_{\text{KL}}(p \parallel \hat{p}) = \sum_{i=1}^P p(x_i) \log \frac{p(x_i)}{\hat{p}(x_i)}, \quad (3.5)$$

which is evaluated over the family of representations generated by varying n . For each dataset realisation, the supervised optimal representation is defined as the one that minimises this divergence, and the corresponding representation size is denoted by n_{KL} .

The comparison carried out in this chapter therefore consists in relating the characteristic optima selected by the **RR** analysis, namely $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$, to the supervised optimum n_{KL} . In this way, we assess whether the representations identified as informative from empirical frequency statistics also correspond to those that best reproduce the underlying probability structure when a reference distribution is available.

3.2.3 Systems and Datasets

The analysis is carried out on synthetic, semi-real, and molecular datasets of increasing structural complexity. In the synthetic and semi-real cases, the reference probability density is known by construction. In the molecular case, where no exact generative distribution is available, a physically motivated reference density is estimated directly from the sampled trajectories. For synthetic and semi-real datasets, the full analysis is repeated over multiple independent realisations; the reported values of $n_{\text{opt}}^{\text{MR}}$, $n_{\text{opt}}^{\text{IT}}$, and n_{KL} correspond to averages over realisations, and uncertainties are given by the associated standard deviations. For molecular dynamics data, the analysis is performed independently on each trajectory, and the resulting values are reported separately.

Unstructured Synthetic Datasets

As a baseline, datasets without latent discrete structure are generated from continuous probability distributions in N dimensions. These systems are used to isolate the effect of dimensionality in the absence of embedded classes or metastable states. In particular, samples are drawn from independent Gaussian, exponential, and beta distributions, as well as from correlated multivariate Gaussian distributions. For all datasets, the estimates of $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$ are obtained using the procedure described in **Appendix A**, which provides both a central value and its associated uncertainty.

For these datasets, the reference probability density $p_{\text{ref}}(x)$ coincides with the generative distribution used to produce the samples. In the simplest case of independent Gaussian components,

$$p_{\text{ref}}(x) = \prod_{\mu=1}^N \mathcal{N}(x_{\mu}; 0, 1), \quad (3.6)$$

with analogous expressions for the other distribution families considered.

Structured Synthetic Datasets

To introduce a controlled form of latent structure, structured synthetic datasets are generated in a space of total dimension $N = 100$, where only a subset of $m < N$ coordinates carries informative signal. Writing each sample as $x = (u, v)$, the variables u define the informative subspace, while the remaining variables v represent an uninformative background.

The informative variables are sampled from a Gaussian mixture model with K components (p_{inf}), while the background variables are drawn as independent Gaussian noise (p_{noise}). The full generative distribution can therefore be written as

$$p_{\text{gen}}(u, v) = p_{\text{inf}}(u) p_{\text{noise}}(v) = \left(\sum_{c=1}^K \pi_c \mathcal{N}(u; \mu_c, \sigma_M^2 I_m) \right) \mathcal{N}(v; 0, \sigma_G^2 I_{N-m}), \quad (3.7)$$

where π_c are the mixture weights, μ_c the component means, σ_M sets the width of the informative components, and σ_G controls the amplitude of the background noise. In the dataset construction, the noise amplitude is chosen such that $\sigma_G \ll \sigma_M$, so that the latent structure is confined to the informative subspace while the remaining dimensions contribute only weakly fluctuating.

The discrete representations are constructed from the full vectors $x \in \mathbb{R}^N$, whereas the reference probability density used in the supervised comparison is defined on the informative variables only. Accordingly, the reference distribution entering the KL divergence is taken to be the informative marginal,

$$p_{\text{ref}}(u) = p_{\text{inf}}(u). \quad (3.8)$$

The KL divergence therefore quantifies the extent to which the representations recover the latent probability structure rather than the uninformative background.

Gaussian Clones of MNIST

As an intermediate setting between fully synthetic and real data, semi-real datasets are constructed as Gaussian clones of MNIST. For each selected digit class, a class-conditional multivariate Gaussian model is estimated from the corresponding MNIST images in the original pixel space. Synthetic samples are then generated from mixtures of these class-conditional Gaussian distributions. The resulting

reference probability density is

$$p_{\text{ref}}(x) = \sum_{c \in C} \pi_c \mathcal{N}(x; \hat{\mu}_c, \hat{\Sigma}_c), \quad (3.9)$$

where C denotes the selected set of digit classes, and $\hat{\mu}_c$ and $\hat{\Sigma}_c$ are the empirical mean and covariance estimated for class c .

Both binary mixtures ($K = 2$) and multi-class mixtures ($K = 5$) are considered, with either equal or unequal class weights. Digit 1 is excluded from the analysis because it systematically produces anomalous **RR** curves, typically displaying a non-standard structure that makes the identification of $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$ non-univocal. A detailed discussion of this effect is reported in [Appendix B](#).

Alanine Dipeptide

As a molecular benchmark, the analysis is applied to **MD** trajectories of alanine dipeptide, a standard test system for conformational analysis. The dataset consists of ten independent trajectories, each 10 ns long. The simulations were performed with OpenMM ([Eastman et al., 2017](#)) on the CUDA platform using the AMBER99SB-ILDN force field ([Lindorff-Larsen et al., 2010](#)) and a GBSA implicit-solvent model ([Onufriev et al., 2000](#)). All trajectories were generated at temperature $T = 300$ K by Langevin dynamics ([Leimkuhler et al., 2013](#)), with friction coefficient $\gamma = 91 \text{ ps}^{-1}$, integration time step $\Delta t = 2 \text{ fs}$, and atomic coordinates saved every 1 ps.

Discrete representations are constructed by clustering molecular configurations in the full configurational space. For the supervised comparison, the reference probability distribution is instead defined in the reduced space of the backbone dihedral angles ϕ and ψ , which provide the standard low-dimensional description of the equilibrium conformational landscape of alanine dipeptide.

For each trajectory, the reference density $p_{\text{ref}}(\phi, \psi)$ is estimated from the sampled dihedral angles by means of a two-dimensional histogram in the (ϕ, ψ) plane. The resulting histogram is normalised and evaluated on the sampled configurations to define the discrete reference probabilities entering the **KL** divergence.

3.3 Results

In this section, the unsupervised optima identified by the **RR** analysis are compared with the supervised optimum obtained by minimising the **KL** divergence from a reference distribution. Throughout the chapter, the comparison is expressed in terms of the ratios between the **KL**-optimal representation size, n_{KL} , and the two characteristic **RR** estimates, namely the maximum-relevance point $n_{\text{opt}}^{\text{MR}}$ and the information-theoretic point $n_{\text{opt}}^{\text{IT}}$, defined by the condition that the local slope of the **RR** curve equals -1 . Ratios close to unity indicate quantitative agreement, whereas

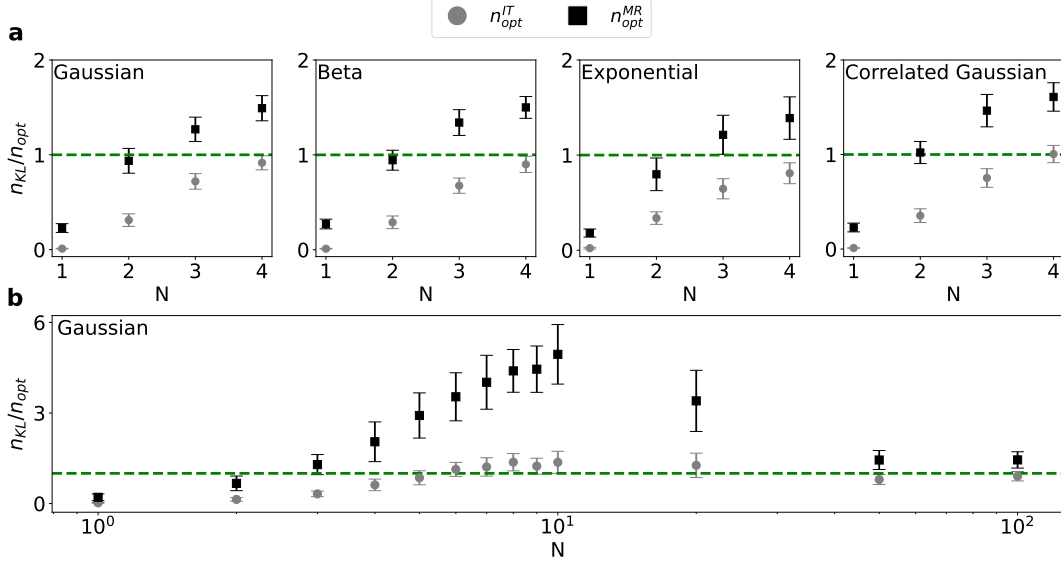


Figure 3.2: Comparison between the *KL*-optimal representation and the two characteristic *RR* estimates for unstructured synthetic datasets. The figure shows the ratio $n_{\text{KL}}/n_{\text{opt}}$ as a function of the dimensionality N , where n_{opt} denotes either the information-theoretic estimate $n_{\text{opt}}^{\text{IT}}$ (gray circles) or the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$ (black squares). The dashed horizontal line marks the value 1, corresponding to perfect agreement with the *KL* optimum. (a) Low-dimensional datasets discretised through N -dimensional histograms. (b) Higher-dimensional iid Gaussian datasets, with representations constructed through agglomerative clustering.

systematic deviations above or below one indicate, respectively, that the *RR* criterion selects representations that are too coarse or too fine relative to the supervised optimum.

3.3.1 Unstructured Synthetic Datasets

As a first benchmark, we considered synthetic datasets generated from continuous distributions without any latent discrete structure. These systems provide the simplest setting in which to isolate the effect of dimensionality, without the additional complication of embedded classes or metastable states.

In the low-dimensional regime, where representations are constructed through N -dimensional histograms (Figure 3.2.a), all distributions display the same qualitative behaviour: for $N = 1$, both *RR* criteria select representations that are finer than the *KL*-optimal one. In this regime, the *RR* framework therefore tends to overestimate the number of discrete states that best reproduces the underlying probability density. As the dimensionality increases, however, this discrepancy decreases rapidly. Already at $N = 2$, the maximum-relevance estimate approaches the supervised optimum, and for larger values of N the *KL*-optimal discretisation systematically falls within the interval delimited by $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$.

The same trend is observed in higher-dimensional iid Gaussian datasets, where the representations are constructed through agglomerative clustering rather than direct histogram discretisation (Figure 3.2.b). Also in this case, the low-dimensional

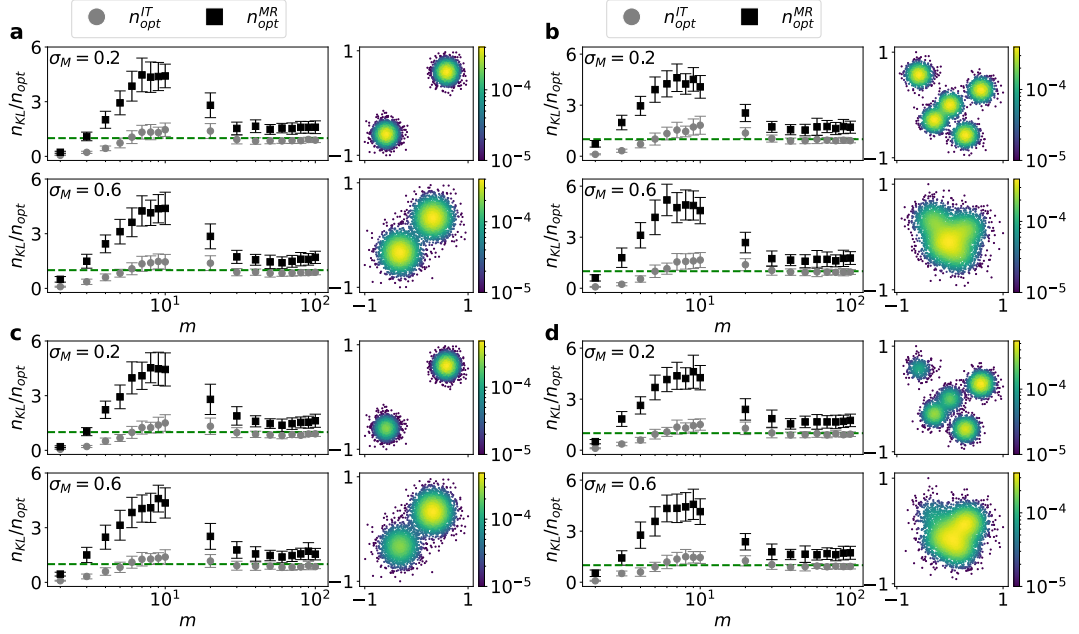


Figure 3.3: Comparison between the KL-optimal representation and the two characteristic RR estimates for structured synthetic datasets. The figure shows the ratio $n_{\text{KL}}/n_{\text{opt}}$ as a function of the number of informative dimensions m , where n_{opt} denotes either the information-theoretic estimate $n_{\text{opt}}^{\text{IT}}$ (gray circles) or the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$ (black squares). The dashed horizontal line marks the value 1, corresponding to perfect agreement with the KL optimum. Panels correspond to Gaussian mixture models with different numbers of components and mixing weights: (a) two components with equal weights, (b) five components with equal weights, (c) two components with unequal weights $[0.66, 0.33]$, and (d) five components with unequal weights $[0.34, 0.27, 0.19, 0.13, 0.07]$. For each case, results are shown for two values of the mixture width, $\sigma_M = 0.2$ and $\sigma_M = 0.6$. The plots on the right illustrate representative datasets in the corresponding informative subspaces.

mismatch progressively disappears as N increases. For intermediate dimensions, the optimality region becomes broader, indicating that the two RR criteria identify a range of informative resolutions rather than a single sharply defined value. Upon further increasing the dimensionality, the two RR estimates move closer to one another, and $n_{\text{opt}}^{\text{IT}}$ approaches the KL optimum.

Taken together, these results establish the baseline observation of the chapter: the mismatch between RR and KL is primarily a low-dimensional effect. Even in the absence of latent structure, increasing dimensionality makes the unsupervised RR criterion progressively more consistent with the supervised notion of optimal fidelity.

3.3.2 Structured Synthetic Datasets

Having isolated the effect of dimensionality in the absence of embedded structure, we next considered synthetic datasets with an explicit latent organisation. These systems were constructed in a space of fixed total dimensionality ($N = 100$), with only a subset of m coordinates carrying informative signal and the remaining ones

acting as uninformative Gaussian background. In the informative subspace, the data were distributed according to Gaussian mixture models with either two or five components, equal or unequal weights. This setting makes it possible to disentangle the role of the nominal dimensionality of the data from that of the *informative dimensionality*, namely the number of coordinates that actually encode the latent probability structure.

As shown in Figure 3.3, the results display the same qualitative trend observed for unstructured datasets, but now as a function of the number of informative dimensions m . When only a very small number of coordinates carries signal, both **RR** criteria tend to select representations finer than the **KL**-optimal one, reproducing the overestimation already observed in low-dimensional continuous systems. As m increases, however, the agreement improves rapidly. In all analysed cases, the **KL**-optimal representation lies within the **RR** interval, and for sufficiently informative datasets it is typically found close to the information-theoretic point $n_{\text{opt}}^{\text{IT}}$.

A further relevant feature concerns the width of the optimality region. For small but increasing values of m , the interval between $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$ initially broadens, indicating that when the latent structure is only partially resolved, the **RR** framework identifies several nearby resolutions as nearly equivalent in descriptive power. As the number of informative dimensions increases further, the interval narrows again, showing that the data support a progressively more sharply defined optimal resolution. In this regime, the **RR** framework not only brackets the supervised optimum, but approaches it with high precision.

These trends are robust with respect to the number of mixture components, the class weights, and the width of the informative clusters, additional combinations of these parameters are reported in Appendix B. The relevant control parameter is therefore not the nominal dimensionality of the ambient space, but the amount of informative structure available in the data. The structured synthetic datasets thus sharpen the interpretation suggested by the previous subsection: the beneficial role of dimensionality is not purely geometric, but depends on the effective dimensionality of the signal.

3.3.3 Gaussian Clones of MNIST

To bridge the gap between fully synthetic models and real-world data, we next considered semi-real datasets constructed as Gaussian clones of MNIST. In this setting, each selected digit class is represented by a class-conditional multivariate Gaussian estimated from the original images, and synthetic samples are drawn from mixtures of these fitted distributions. This procedure preserves important aspects of the covariance structure of the original handwritten-digit data while still providing an explicit reference probability model for the **KL** comparison. The analysis was carried out for binary and five-class mixtures, with either equal or unequal class weights, considering all admissible combinations of digit classes except those involving digit

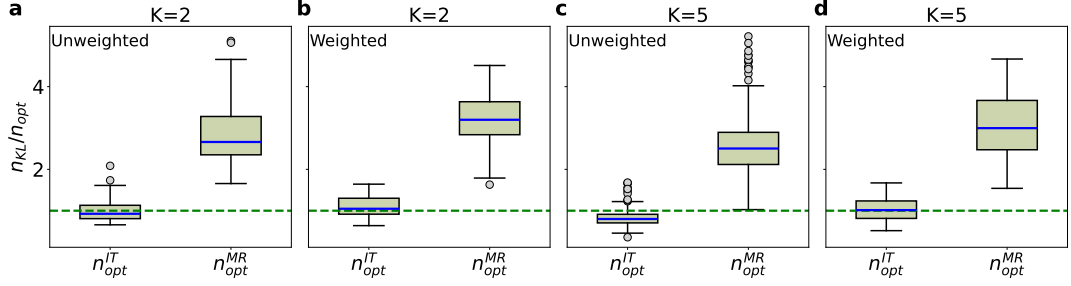


Figure 3.4: Comparison between the *KL*-optimal representation and the two characteristic *RR* estimates for Gaussian clones of MNIST. The figure shows the distribution of the ratio $n_{\text{KL}}/n_{\text{opt}}$ for different class mixtures, where n_{opt} denotes either the information-theoretic estimate $n_{\text{opt}}^{\text{IT}}$ or the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$. The dashed horizontal line marks the value 1, corresponding to perfect agreement with the *KL* optimum. Panels correspond to mixtures with different numbers of classes and different class weights: (a) two-class mixtures with equal weights, (b) two-class mixtures with unequal weights ([0.66, 0.33]), (c) five-class mixtures with equal weights, and (d) five-class mixtures with unequal weights ([0.34, 0.27, 0.19, 0.13, 0.07]).

1, whose anomalous *RR* behaviour is discussed separately in the [Appendix B](#).

As shown in [Figure 3.4](#), the information-theoretic point $n_{\text{opt}}^{\text{IT}}$ display the best agreement with the *KL*-optimal discretisation across all tested mixture configurations. The corresponding distributions of the ratio $n_{\text{KL}}/n_{\text{opt}}^{\text{IT}}$ are centred close to one, indicating that the -1 slope criterion provides, on average, a nearly unbiased estimate of the supervised optimum. By contrast, the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$ systematically selects coarser representations, yielding values of $n_{\text{KL}}/n_{\text{opt}}^{\text{MR}}$ larger than one.

This distinction between the two *RR* criteria is consistent with their respective roles. The maximum-relevance point identifies the representation at which occupancy heterogeneity is maximal, whereas the information-theoretic point marks a more permissive threshold, at which additional resolution can still be introduced without a net informational loss. In a structured and high-dimensional dataset such as MNIST, this second criterion is the one that most closely matches the supervised optimum. In other words, once the data contain sufficiently rich internal organisation, the best representation is generally not the one at which relevance is maximal, but the one slightly beyond it, where further descriptive detail remains statistically meaningful.

Another relevant observation is that this behaviour is stable across different numbers of mixture components and across both equal and unequal class weights. The close correspondence between $n_{\text{opt}}^{\text{IT}}$ and n_{KL} is therefore not a special feature of a particular class combination, but a robust property of this semi-real setting. The Gaussian-clone construction thus provides a strong intermediate validation of main claim: when the data are both high-dimensional and sufficiently structured, the information-theoretic *RR* criterion becomes a reliable unsupervised proxy for the supervised optimum.

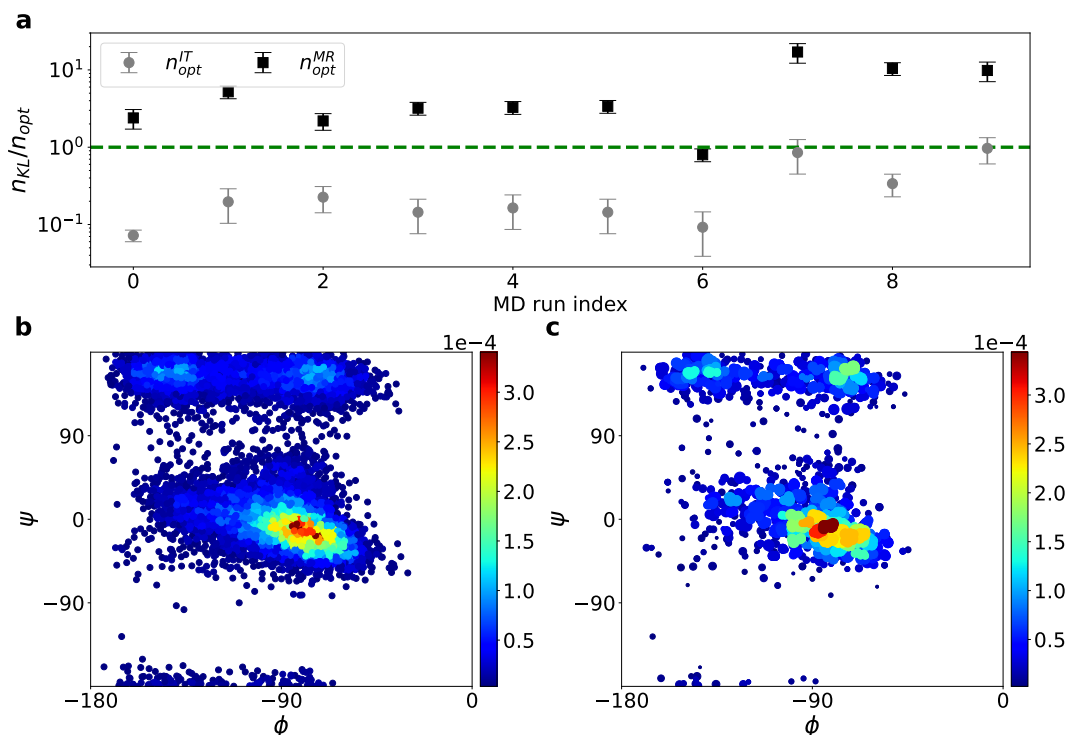


Figure 3.5: Comparison between the KL -optimal representation and the two characteristic RR estimates for alanine dipeptide. The figure shows the ratio $n_{\text{KL}}/n_{\text{opt}}$ for 10 independent molecular dynamics trajectories, where n_{opt} denotes either the information-theoretic estimate $n_{\text{opt}}^{\text{IT}}$ (gray circles) or the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$ (black squares). The dashed horizontal line marks the value 1, corresponding to perfect agreement with the KL optimum. (a) Values of $n_{\text{KL}}/n_{\text{opt}}$ as a function of the MD run index. (b) Equilibrium probability density in the (ϕ, ψ) dihedral-angle space estimated from the molecular dynamics trajectory. (c) Coarse-grained representation of the same conformational landscape obtained from the clustering at $n_{\text{opt}}^{\text{IT}}$.

3.3.4 Alanine Dipeptide

As a final test, we considered a molecular dataset consisting of ten independent molecular dynamics trajectories of alanine dipeptide. Unlike the previous cases, no exact generative distribution is available here. The supervised reference is therefore defined in a physically motivated way, by estimating the equilibrium probability density in the space of the backbone dihedral angles ϕ and ψ , which provide the standard low-dimensional description of the conformational landscape of this molecule. The discrete representations entering the RR analysis, however, are built in the full configurational space through RMSD-based clustering. The comparison, therefore, probes whether the discretisations selected from full atomic configurations recover the physically relevant equilibrium structure captured in dihedral-angle space.

As shown in Figure 3.5, the KL -optimal number of clusters lies within the RR optimality region for all analysed trajectories. This confirms that the unsupervised criterion remains consistent with the supervised one even in an empirical molecular system, where the reference distribution is not known by construction but estimated

from finite trajectories. At the same time, the alanine dipeptide does not display the same sharp preference for one **RR** criterion over the other that was observed in the structured synthetic and MNIST-based datasets: in some trajectories the **KL** optimum is closer to $n_{\text{opt}}^{\text{MR}}$, whereas in others it is closer to $n_{\text{opt}}^{\text{IT}}$.

This broader variability is expected. In the molecular case, the supervised reference is an empirical estimate obtained from the same finite trajectory and projected onto a reduced set of collective variables, while clustering is performed in the full configurational space. The two descriptions are therefore not expected to coincide exactly, even in the absence of noise. What matters is that the **RR** framework consistently restricts the plausible optimal discretisation to a narrow interval, and that this interval remains stable across independent trajectories.

A qualitative inspection of the corresponding representations further supports this interpretation. In [Figure 3.5b](#), each point corresponds to an atomistic frame projected onto the (ϕ, ψ) plane, whereas in [Figure 3.5c](#) the coarse-grained description obtained at $n_{\text{opt}}^{\text{IT}}$ is shown by representing each cluster through its centroid, with marker size proportional to the cluster population. This representation reproduces the main large-scale features of the conformational landscape. The principal basins and their relative organisation are preserved, even though the representation is obtained without using dihedral information during clustering. This shows that the discretisations selected by the **RR** analysis are not only statistically consistent with the **KL** benchmark, but also physically meaningful from the perspective of the standard conformational coordinates of alanine dipeptide.

3.4 Conclusions

The first main result of this chapter is that the unsupervised **RR** criterion and the supervised **KL** benchmark show a substantial agreement in the selection of informative representations. Across the different datasets analysed here, the **KL**-optimal representation is generally found close to, or within, the **RR** optimality region delimited by $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$. A second central result is that this agreement becomes progressively stronger as the data become more informative. In low-dimensional or weakly structured settings, the two **RR** optima tend to select representations that are more refined than the **KL**-optimal one. This mismatch, however, decreases when either the dimensionality of the data or the number of informative degrees of freedom is increased. In the most structured high-dimensional cases, $n_{\text{opt}}^{\text{IT}}$ becomes a very good approximation of n_{KL} .

This result does not imply that increasing dimensionality automatically makes representation selection easier. Rather, it shows that when the data possess a sufficiently rich internal structure, the **RR** trade-off becomes increasingly aligned with the criterion based on fidelity to the reference probability distribution. The structured synthetic datasets make this point particularly clearly: what drives the agreement is not the ambient dimensionality by itself, but the number of coordinates that

actually carry informative signal. In this sense, the beneficial effect is tied to the effective dimensionality of the organisation encoded in the data.

The analysis also suggests that the most robust output of the **RR** framework is not always a single preferred number of states, but rather an optimality region. This point emerges most clearly in the alanine dipeptide data, where the **KL**-optimal discretisation is consistently found within the interval bounded by $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$, but does not show a systematic proximity to one specific boundary across all trajectories. In this case, the interval provides a more stable and meaningful description than either characteristic estimate taken in isolation. By contrast, in the synthetic and semi-real datasets, and in particular in the Gaussian clones of MNIST, $n_{\text{opt}}^{\text{IT}}$ is generally the closest estimate to the supervised optimum. The two characteristic points therefore play distinct roles: $n_{\text{opt}}^{\text{MR}}$ marks the lower boundary of the **RR** optimality region, while $n_{\text{opt}}^{\text{IT}}$ often provides the best quantitative approximation to n_{KL} when the structure of the data is sufficiently rich.

At the same time, the present analysis helps clarify the scope of what **RR** can and cannot assess. Since the framework evaluates representations through the distribution of their occupancies, its criterion is entirely determined by the balance between resolution and relevance encoded in that distribution. It is therefore informative only when applied to a family of candidate representations constructed according to a meaningful prescription, such as a geometric partition, a clustering procedure based on a sensible distance, or a physically motivated measure of similarity. Within such a family, **RR** can identify the level of resolution that is most informative. By contrast, it does not directly assess the internal composition of the groups themselves. Different partitions sharing the same occupancy distribution are indistinguishable from the point of view of **RR**, even if one preserves the structure of the data while another does not. The method should thus be understood as a criterion for selecting the resolution within a prescribed family of representations, rather than as a stand-alone criterion for validating an arbitrary grouping of the data. A more quantitative discussion of this point is provided in [Appendix B](#).

The numerical evidence collected in this chapter strongly suggests that the correspondence between **RR** and **KL** is not accidental, at least for broad classes of datasets. A natural extension of the present work would therefore be the derivation of an analytical relation connecting the **KL** optimum to the structure of the **RR** curve. Such a result would clarify under which assumptions the **KL**-optimal representation is expected to fall within the **RR** optimality region, and under which conditions it should approach one of its two boundaries. This would place the empirical trends reported here on firmer conceptual grounds.

In conclusion, this chapter shows that the **RR** framework provides a consistent unsupervised indication of representation size when the candidate representations are built in a meaningful way. The central result is not an exact equivalence between **RR** and **KL**, but the observation that, as the informative content of the data increases, the representation selected by **KL** moves into the **RR** optimality region

and often approaches $n_{\text{opt}}^{\text{IT}}$. In this sense, the chapter provides a bottom-up validation of the **RR** criterion in the supervised settings where such a comparison is possible, and supports its use as a practical guide for selecting informative low-resolution representations in structured high-dimensional systems.

Take-home message

Selecting the resolution of a discrete representation is difficult when the underlying probability distribution is unknown. In this chapter, the **Resolution-Relevance** framework was tested against a supervised benchmark based on the **Kullback-Leibler** divergence, in settings where the reference distribution was known. The results showed that the representation selected by the supervised criterion was generally contained within the **Resolution-Relevance** optimality region, delimited by the maximum-relevance point and the information-theoretic point. The agreement improved when the data contained richer and higher-dimensional informative structure, supporting the use of **Resolution-Relevance** as an unsupervised guide for choosing the resolution of meaningful candidate representations.

4

Configurational Coarse-Graining

4.1 Introduction

Molecular Dynamics (MD) simulations provide an increasingly detailed view of the configurational space explored by biomolecules, generating trajectories that contain a large amount of structural information over many length and time scales. Advances in force fields, integration schemes, and hardware now make it possible to investigate systems ranging from membrane assemblies and large protein complexes to whole viruses and even organelle-scale models, producing trajectories that contain structural information across many length and time scales (Gapsys et al., 2024; Beck et al., 2024; Mathur et al., 2024). The central difficulty consequently shifts from sampling configurations to extract, from them, a compact description that remains physically meaningful (Noid, 2013; Noid et al., 2024). In practice, this requires grouping similar conformations into a smaller number of effective states, so as to separate robust collective organisation from microscopic detail and thermal fluctuations. This problem can be formulated as a configurational **CG** task, in which the trajectory is replaced by a partition of the sampled conformational space.

A natural way to construct such partitions is through clustering (Abramyan et al., 2016). In the case of protein trajectories, clustering methods typically rely on a structural notion of similarity, such as the root mean square deviation between frames after roto-translational alignment, and use it to group nearby configurations into common states (Grossfield et al., 2018; Papaleo et al., 2009). However, the result of this procedure depends not only on the underlying data, but also on the

This chapter is based on material originally published in M. Mele, R. Covino, and R. Potestio, *Information-theoretical measures identify accurate low-resolution representations of protein configurational space*, *Soft Matter* 18(37), 7064–7074 (2022). The material has been revised and adapted for consistency with the structure and notation of this thesis, and is reproduced in accordance with the terms of the applicable licence. The published version is available at: <https://doi.org/10.1039/D2SM00636G>.

specific clustering strategy used to organise them. Different clustering linkage criteria can produce markedly different low-resolution descriptions even when they act on the same trajectory and the same distance matrix (Murtagh et al., 2012; Tokuda et al., 2022). As a consequence, the problem is not only to determine how many states should be retained, but also to understand which clustering procedures generate partitions that preserve the informative structure of the original configurational space.

The present analysis sets into the broader path developed in the previous chapters. After introducing the **Resolution-Relevance (RR)** framework as a way to quantify the balance between descriptive detail and statistically reliable structure (Section 2.3), the thesis has shown that this approach can provide a consistent unsupervised indication of the appropriate representation scale (Chapter 3), provided that the candidate descriptions are constructed in a meaningful way. At the same time, that analysis also showed that **RR** cannot by itself validate an arbitrary partition, since partitions with the same occupancy distribution are equivalent from the point of view of relevance and resolution. The problem addressed in this chapter follows directly from this observation: rather than asking whether **RR** can validate a partition in absolute terms, the question is whether it can discriminate between different physically motivated clustering procedures applied to protein configurational data.

To address this problem, this chapter analyses the configurational **CG** of protein **MD** trajectories through agglomerative clustering (Murtagh et al., 2012), comparing different linkage criteria and different structural representations. The analysis is first performed on a dataset of structurally distinct proteins (Tarenzi et al., 2022) and then on adenylate kinase as a detailed case study (Beckstein et al., 2009), complemented by results on a humanised IgG4 antibody as an additional validation on a larger and more complex biomolecular system (Tarenzi et al., 2021). The aim is twofold: to determine whether **RR**-based quantities can identify those clustering procedures that yield more informative low-resolution representations of the trajectory, and to clarify what these results reveal about the organisation of protein configurational space itself. As will be shown, the clustering procedures favoured by the **RR** analysis are also those that best preserve the geometry of the high-resolution reference, supporting a hierarchical and approximately ultrametric view of the conformational landscape.

4.2 Materials and Methods

The analysis presented in this chapter compares different clustering procedures for the configurational **CG** of protein **MD** trajectories. For each system, a family of partitions is generated by agglomerative clustering, varying both the structural representation used to compute configurational distances and the linkage criterion used to define inter-cluster similarity. The resulting partitions are then evaluated from two

complementary view points: their position in the **RR** plane, quantified through relevance and **Multiscale Relevance (MSR)**, and their ability to preserve the structural organisation of the reference high-resolution trajectory, assessed through diffusion maps and covariance decomposition.

4.2.1 Systems and Simulation Data

The analysis was carried out at two complementary levels. First, an exploratory dataset of structurally distinct proteins was used to identify general trends across different biomolecular systems. The dataset was selected from the dynamical classification introduced in [Tarenzi et al., 2022](#), a set of 107 proteins, covering several folds and structure classes, was grouped according to the similarity of their essential dynamics. The 12 proteins considered in the present work correspond to the centroids of the resulting clusters and therefore provide a set of dynamically distinct representatives.

For the detailed analysis, two additional systems were considered. Adenylate kinase was used as the main case study, as it provides a well-established example of functionally relevant conformational variability on accessible simulation timescales ([Beckstein et al., 2009](#)). In addition, a humanised IgG4 monoclonal antibody was included as a larger and structurally more complex validation system ([Tarenzi et al., 2021](#)); the corresponding results are reported in [Appendix C](#).

All trajectories were generated with GROMACS 2018 using the AMBER99SB-ILDN force field in explicit TIP3P water, with Na^+ and Cl^- ions at a concentration of 0.15 M to neutralise the system and reproduce physiological ionic strength ([Bekker et al., 1993](#); [Abraham et al., 2015](#); [Lindorff-Larsen et al., 2010](#)). After energy minimisation, NVT and NPT equilibration were performed using the velocity-rescale thermostat and the Parrinello-Rahman barostat ([Bussi et al., 2007](#); [Parrinello et al., 1981](#)). Short-range van der Waals and Coulomb interactions were treated with a cut-off, long-range electrostatics were handled with the Particle Mesh Ewald method, and hydrogen-containing bonds were constrained with LINCS, allowing the use of a 2 fs integration time step ([Hockney et al., 2021](#); [Hess et al., 1997](#)). The systems in the 12-protein dataset were simulated for 300 ns each, while adenylate kinase and the humanised IgG4 antibody were simulated for 800 ns and 2 ms, respectively.

Configurational Representations and Clustering

For each protein, configurational similarity was quantified from the trajectory by computing the pairwise RMSD matrix after removal of global translations and rotations. To probe the effect of structural resolution on the resulting partitions, this procedure was repeated for three atom selections: all atoms, C_α atoms, and C_β atoms. The all-atom representation retains the full structural detail of the protein, whereas the C_α and C_β descriptions provide residue-level projections with one effective site per amino acid. Each system is therefore associated with three config-

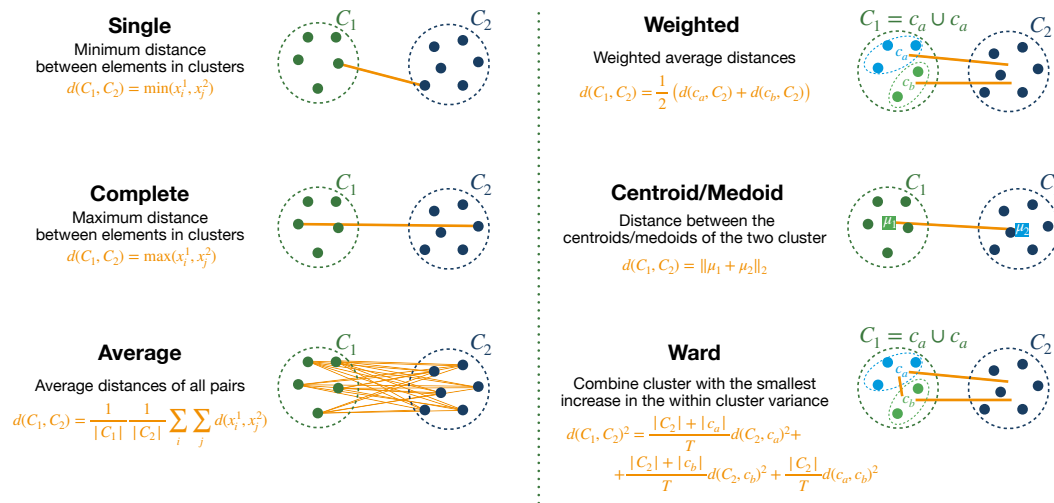


Figure 4.1: Schematic illustration of the linkage criteria adopted in the agglomerative clustering procedure. Different definitions of inter-cluster distance produce different dendrograms and, consequently, different configurational partitions.

urational distance matrices corresponding to different structural representations of the same trajectory.

Starting from each RMSD matrix, configurational partitions were generated by agglomerative clustering. Seven linkage criteria were considered, each defining the distance between clusters in a different way: single, complete, average, weighted, centroid, median, and Ward linkage. Their operational differences are schematically illustrated in Figure 4.1. For each atom selection and each linkage rule, a dendrogram was constructed and cut at different levels by varying the number of clusters K from 1 to the total number of frames M . In this way, each cut of the dendrogram defines a CG representation of the trajectory in which the original configurations are grouped into K configurational states. Overall, this procedure yields 21 RR curves for each system, corresponding to the combination of 3 structural representations and 7 clustering protocols.

For every partition, resolution and relevance were computed from the frequency distribution of the resulting clusters. In particular, if k_s denotes the number of trajectory frames assigned to cluster s , the resolution and relevance are computed from the distribution of cluster frequencies $\{k_s\}$ according to Equation 2.26 and Equation 2.27. In the present chapter, these quantities are used as operational observables to compare the partitions obtained from different clustering procedures. Their overall informativeness across different scales is summarised by the Multiscale Relevance (MSR), defined as the area under the corresponding RR curve. This makes it possible to compare clustering protocols not at a single value of K , but in terms of their ability to generate informative representations over a broad range of CG levels.

As a null model, we also considered random clustering. For a fixed number of clusters K , random label vectors were generated by assigning each of the M trajectory frames to an integer in the range $\{1, \dots, K\}$. For each value of K , we sampled

10^4 random partitions and computed the corresponding resolution and relevance values, which were then averaged to obtain the points of the random **RR** curve. Repeating this procedure for all K from 1 to M yields a baseline against which the agglomerative clustering results can be assessed. The **MSR** associated with this random curve (MSR_R) was then used to normalise the **MSR** values obtained from the hierarchical clustering protocols

$$\overline{MSR} = \frac{MSR - MSR_R}{MSR_R}. \quad (4.1)$$

4.2.2 Diffusion Maps and Covariance Decomposition

To characterise the structural properties of the partitions obtained from clustering, we employed two complementary analyses: diffusion maps and covariance decomposition.

Diffusion maps are a non-linear dimensionality-reduction technique that represents high-dimensional data through the leading modes of a diffusion process defined on a similarity graph. Here, they were used as a geometric validation tool to assess whether the low-resolution (LR) partitions preserve the large-scale organisation of the original configurational space. In this framework, the configurations are represented as the nodes of a weighted graph whose connectivity is determined by the chosen distance matrix through a kernel function. A diffusion process is then defined on this graph, and its leading eigenmodes provide a low-dimensional embedding in which points that are strongly connected in the original space remain close, whereas weakly connected or distant points are mapped farther apart (De la Porte et al., 2008; Lafon et al., 2006; Nadler et al., 2006). The comparison was therefore performed between the manifold reconstructed from the full trajectory and that obtained from the corresponding LR representations. In the high-resolution (HR) case, each point of the manifold corresponds to a trajectory frame and the pairwise distance is the same RMSD matrix used for clustering. In the LR case, each point corresponds to a cluster centroid, and the distance between points is defined by the linkage measure adopted to construct the partition. The diffusion-map kernel was constructed using as threshold the quantile of order 0.1 of the corresponding distance distribution, and the resulting diffusion coordinates were rescaled so that all points were contained in the square $[-1, 1]^2$. The similarity between the HR and LR manifolds was then quantified through the Pearson correlation coefficient (PCC) between the first two diffusion coordinates obtained in the two representations.

As a complementary structural diagnostic, the covariance matrix of the C_β positions along the trajectory was decomposed into intra- and inter-cluster contributions (Pontiggia et al., 2008; Kitao et al., 1998). The full covariance matrix is written as

$$C_{ij} = C_{ij}^{\text{intra}} + C_{ij}^{\text{inter}}, \quad (4.2)$$

with

$$C_{ij}^{\text{intra}} = \sum_l \omega_l \left\langle (r_i - \langle r_i \rangle_l) (r_j - \langle r_j \rangle_l) \right\rangle_l, \quad (4.3)$$

and

$$C_{ij}^{\text{inter}} = \sum_l \omega_l (\langle r_i \rangle_l - \langle r_i \rangle) (\langle r_j \rangle_l - \langle r_j \rangle). \quad (4.4)$$

Here, l runs over the clusters, ω_l is the statistical weight of cluster l , namely the fraction of simulation time spent in that state, and $\langle \cdot \rangle_l$ denotes the average restricted to the conformations belonging to cluster l . In this decomposition, the intra-cluster term measures the structural fluctuations within each state, whereas the inter-cluster term accounts for the structural differences between cluster centroids. In practice, this comparison was performed through the traces of the two covariance components as a function of the number of clusters.

4.3 Results

The results presented in this section address two complementary aspects of the configurational CG problem. First, the 11-protein dataset is used to compare the informativeness of different clustering protocols across structurally distinct systems and different levels of structural resolution. Second, adenylate kinase is employed as a detailed case study to assess whether the partitions favoured by the RR analysis also preserve the geometric and structural organisation of the underlying configurational space. Corresponding results for the humanised IgG4 antibody are reported in Appendix C.

4.3.1 12-Protein Dataset

We first investigate the effect of the linkage criterion on the RR curves obtained by clustering protein configurations sampled from MD trajectories. Figure 4.2 shows that, even when the same trajectory is considered, the resulting curves depend markedly on both the clustering protocol and the structural representation used to define configurational similarity. In particular, changing the atom selection can substantially modify the position of the curve in the RR plane, indicating that the informativeness of the resulting partition depends not only on the sampled trajectory, but also on the level of structural detail retained in the description.

To compare the different protocols across proteins and representations, we quantify the overall informativeness of each curve through the MSR, namely the area under the RR curve. The corresponding results are summarised in Figure 4.3, where the relative MSR with respect to the random reference is plotted against the mean RMSD of the trajectory. This representation makes it possible to compare the performance of the seven linkage criteria over the full dataset and to relate it to the structural mobility explored by each system.

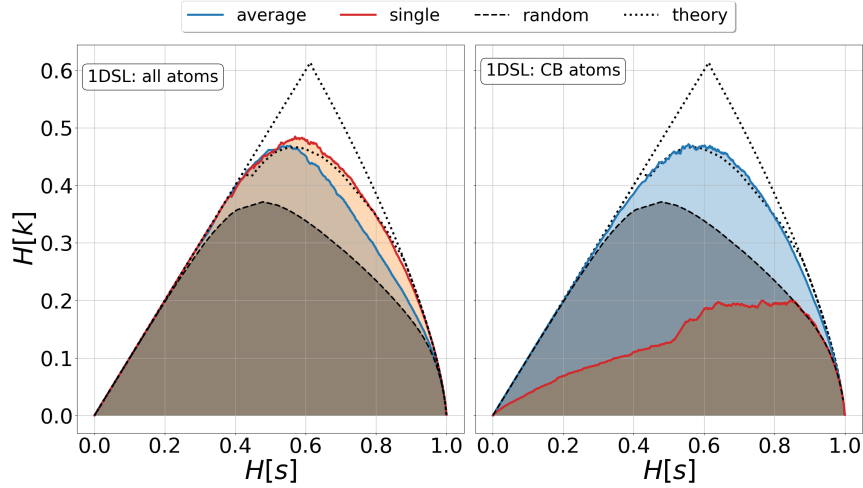


Figure 4.2: *RR* curves obtained partitioning the simulation data of protein 1DSL with two clustering protocols, average linkage and single linkage. The two panels differ by the atom selection adopted: all atoms on the left and C_β atoms on the right. In both panels the random curve is also present, obtained by randomly partitioning the structures into groups. Each $(H[s], H[k])$ point corresponds to a fixed number of clusters in which the frames of the whole trajectory are grouped. The theoretical highest- $H[k]$ for given $H[s]$ are also plotted (black dotted line); specifically, the dotted lines show the upper and lower bounds to the theoretical maximum.

Two main behaviours emerge. Average, complete, weighted, and Ward linkage systematically yield positive relative **MSR** values and remain close to the theoretical maximum, independently of the specific protein and of the structural representation adopted. By contrast, centroid and median linkage display a stronger dependence on the dataset and on the representation: their performance correlates with the mean RMSD and deteriorates when the structural description is reduced to residue-level representations. This indicates that these methods are less robust, since their ability to identify informative **CG** descriptions depends more strongly on the specific system under examination.

The single linkage criterion shows a peculiar pattern and deserves separate discussion. Here, the distance between clusters is defined as the minimum pairwise distance between their elements. Hence, in the all-atom case, where the RMSD matrix elements are broadly distributed, this algorithm is still able to form clusters with heterogeneous populations. Conversely, the C_α and C_β selections implement a **CG** representation that “blurs” structural differences from the outset. As a direct consequence of the single-linkage definition, clusters can then be merged whenever even just one pair of frames is close, leading to the grouping of also relatively distant configurations into the same cluster. This is the well-known chaining effect, which in this case yields a rather uninformative representation of the system.

It can be shown (Milligan, 1979) that some of the hierarchical clustering algorithms induce a monotonic hierarchy, *i.e.*, the values in the inter-cluster distance matrix increase monotonically during agglomerative clustering. Algorithms that induce a monotonic hierarchy lead to an ultrametric in the cluster space (Johnson, 1967): this implies that the metric distance satisfies an inequality stronger than the

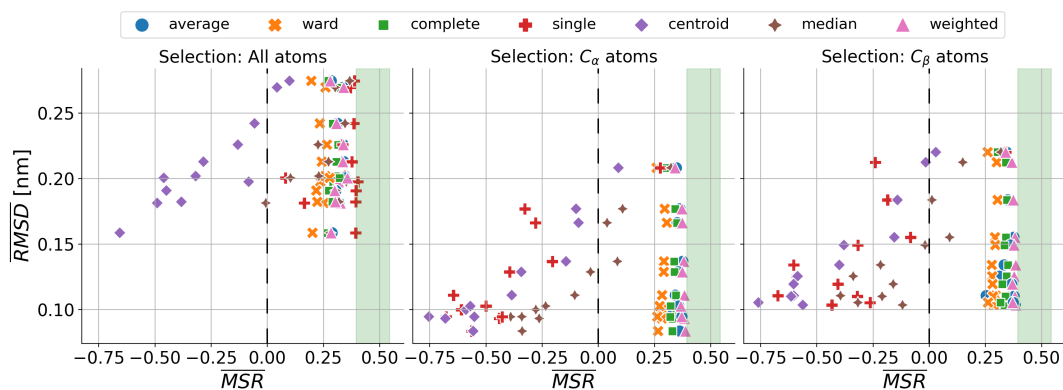


Figure 4.3: Performance of clustering algorithms, quantified by the relative MSR ($\overline{MSR}$), related to the protein mobility, given by the mean value of the $RMSD$ matrix ($\overline{RMSD}$). Each point corresponds to one of the 12 proteins of the dataset. The coloured region highlights the range of MSR values associated to the RR curves falling between the theoretical limits to the curve of maximal relevance. The three panels differ by the atom selection adopted for clustering: all atoms on the left, C_α atoms in the middle, and C_β atoms on the right. The colours and the markers identify the linkage criterion used in the clustering procedure as shown in legend.

triangular one (Jardine et al., 1968; Fushing et al., 2013). In our analysis, it turns out that clustering protocols that satisfy these qualities coincide with those showing a consistently positive MSR ; the only exception to this trend is single linkage, which, although inducing an ultrametric in cluster space, still shows negative MSR values when CG representations of the system are employed. In this case, however, the clustering protocol is severely limited by the chaining effect, plausibly producing uninformative partitions of the system and consequently obtaining a lower or comparable MSR value with respect to the random case.

Taken together, these results suggest that protein structures sampled in the course of a molecular dynamics simulation populate the configurational space according to an ultrametric structure, which is consistent with the self-similar organisation of the free energy landscape observed in previous works (Pontiggia et al., 2007; Volkhardt et al., 2022; Pandya et al., 2018; Henzler-Wildman et al., 2007); additionally, the MSR appears to be capable of capturing, in a parameter-free and unbiased manner, the effectiveness of a clustering method in finding informative representations of a biomolecule’s configurational variability at different scales of resolution, in that MSR correlates with the method’s capacity to preserve the ultrametric structure of the reference configurational space.

4.3.2 Adenylate Kinase

We next examined in greater detail how the linkage criterion affects the informativeness of the resulting low-resolution (LR) representation of a protein conformational space. As a case study, we considered adenylate kinase (Beckstein et al., 2009; Ballard et al., 2017). Starting from configurations extracted from an 800 ns simulation and reduced to the positions of the C_β atoms only, we constructed clusterings us-

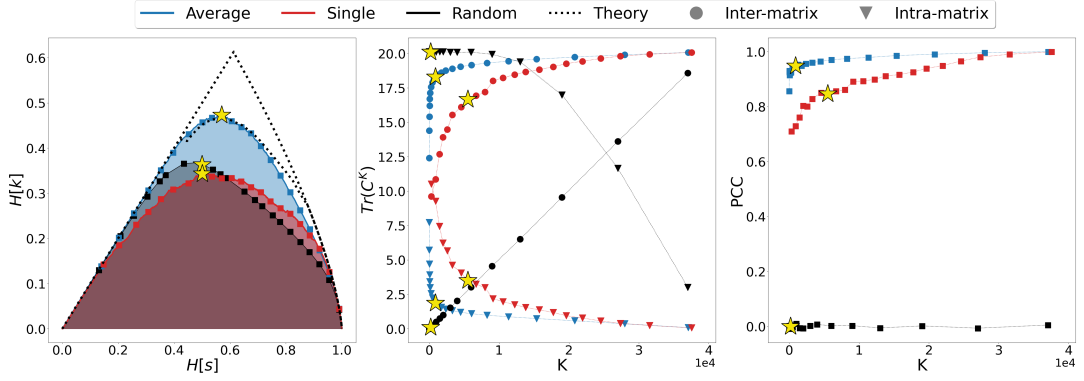


Figure 4.4: Comparison of the different LR representations obtained from clustering the MD trajectory of adenylate kinase using the average linkage (blue), single linkage (red), and random clustering (black). The system was analysed using a coarse representation in which only C_β atoms are considered. Left panel: RR curves drawn by the different protocols; the square points mark the 22 low-resolution representations under examination. The theoretical limits of the maximal relevance curve are also shown as dotted lines for comparison. Middle panel: traces of the inter-state (circles) and intra-state (triangles) correlation matrix plotted against the number of clusters K . Right panel: Pearson correlation coefficient (PCC) between the first two DCs in the HR and LR representation, plotted against the number of clusters K . The yellow star in each of the graphs indicates the representation that maximises the relevance for the corresponding clustering method ($K = \{901, 5511, 200\}$ for average, single or random protocol respectively).

ing single, average, and random linkage at different numbers of clusters. These correspond to 22 distinct resolution values in the interval $[0, 1]$: the spacing in resolution is approximately 0.06 for the first seven representations and approximately 0.04 for the remaining ones. For each LR partition, we computed the quantities reported in Figure 4.4: the RR plots (left panel), the traces of the inter-state and intra-state covariance matrices (middle panel), and the Pearson correlation coefficient (PCC) between the first two diffusion coordinates in the high-resolution (HR) and LR spaces (right panel). The latter two quantities are shown as functions of the number of clusters K used in the representation.

At fixed resolution, the LR representations obtained with average linkage are both simpler and more informative than those obtained with single linkage: they achieve higher relevance values with a smaller number of clusters K . By contrast, single linkage yields a relevance curve that lies very close to that of the random partition, whereas the average-linkage curve approaches much more closely the lower bound to the theoretical maximum. This ranking is confirmed by both the covariance traces and the PCC. In particular, already for a small number of clusters ($K \sim 10$), average linkage identifies LR representations in which the inter-cluster contribution is substantially larger than the intra-cluster one. Single linkage, instead, produces partitions in which these two contributions are of the same order, or even reversed, with the intra-state term larger than the inter-state one.

These results indicate that LR representations produced by single-linkage clustering do not adequately preserve the information contained in the data: they discard an amount of information comparable to, or even greater than, what they

retain. More generally, for average linkage the inter-state contribution to the covariance is always significantly larger, and the intra-state contribution significantly smaller, than for single linkage. The two methods become comparable only when more than half of the original frames are retained. Dimensionality-reduction results support the same conclusion: consistently with the trend of the **MSR**, the **PCC** is systematically higher for average linkage than for single linkage, showing that the former produces more informative LR representations.

In all three panels of **Figure 4.4**, yellow stars mark the representations that maximise the relevance for each clustering method. Interestingly, in both the covariance plots and the **PCC** plots, these points are located at the elbow of the corresponding curve. Additional analyses supporting this interpretation, in particular concerning the location of the point with slope $\mu = 1$ and the behaviour of the inter-cluster distance as a function of cluster number, are provided in **Appendix C**.

Overall, this behaviour suggests that increasing the resolution beyond this point leads to a rise in model complexity that is no longer compensated by a corresponding gain in information. In other words, the trade-off between complexity and informativeness shifts in favour of the former, consistently with the interpretation of relevance as a measure of useful information content. The **RR** framework therefore makes it possible to identify an optimal set of representative configurations for the dataset. Any observable can then be evaluated on these configurations, provided that their statistical weights, proportional to the size of the corresponding cluster, are properly taken into account. This leads to a substantial reduction in computational cost: assuming linear scaling with dataset size, the gain is proportional to the ratio between the number of original data points and the number of clusters. For the representation identified by average linkage, this ratio is approximately 50.

To further assess the statistical significance of these findings, we compared the results for average and single linkage with those obtained from random clustering. The relevance curve for random clustering is very close to that of single linkage, and the corresponding **MSR** values, $MSR_S = 0.235$ and $MSR_R = 0.233$, differ only at the third decimal place. Despite this similarity, the actual usefulness of the partitions obtained with an information-driven procedure is vastly greater than that of random clustering. For the latter, the trace of the inter-cluster covariance matrix remains smaller than the intra-cluster one until representations retaining about 2/3 of the original frames are considered, and the **PCC** between the diffusion coordinates of the reference and random partitions is nearly zero at all resolution levels. These observations reinforce the idea that relevance alone cannot be regarded as an absolute measure of the informativeness of a given LR representation. However, when combined with an appropriate clustering criterion, it becomes a highly effective quantity for identifying protocols that maximise the emergence of useful information.

Finally, we examined in detail the three representations that maximise the relevance for each of the clustering methods considered. In the right-hand side of

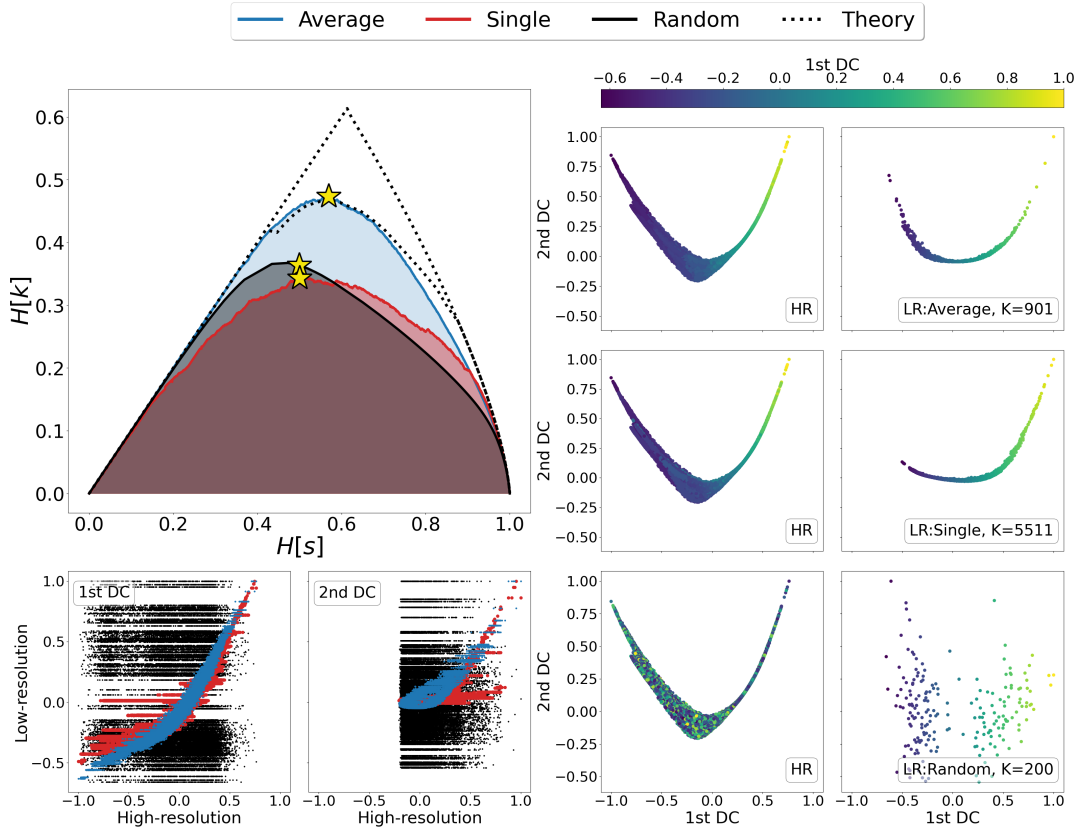


Figure 4.5: Detailed analysis of the highest-relevance partitions of adenylate kinase trajectory data obtained with single linkage, average linkage, and random clustering. Upper left-hand panel: *RR* curves obtained from clustering the MD trajectory of adenylate kinase via average linkage (blue), single linkage (red), and random clustering (black). The system was analysed using a *CG* representation in which only the C_{β} atoms are considered. The yellow stars indicate the position, on the corresponding curve, of the low-resolution representations maximising the relevance obtained by partitioning the system into a K clusters. The dotted lines show the theoretical upper and lower bounds to the maximal relevance curve. Right panel: trajectory frames or cluster points projected onto the space spanned by the first two diffusion coordinates obtained by diffusion maps. The panel's left column shows the low-dimensional manifold resulting from high-resolution (HR) representation where each point is a frame of the MD simulation; the right column shows the two-dimensional manifold resulting from the low-resolution (LR) representation, where each point is the centroid of a cluster. All HR diffusion maps were calculated using the RMSD between configurations as a distance. For the LR manifold, instead, the distance between points is defined by the linkage measure used to construct the corresponding partition for the information-driven clustering protocols, namely average linkage (first row) and single linkage (second row); for random clustering, it is defined as the RMSD between cluster centroids (third row). In both HR and LR spaces, the points are coloured according to the value taken by the first diffusion coordinate in the LR space as reported in the colour bar at the top. Lower left-hand panel: scatter plot of the first (left) and second (right) diffusion coordinates of the HR space plotted against the corresponding coordinates of the LR space; in each graph, we report the points of the representations obtained through random clustering (black), single linkage (red), and average linkage (blue). Note that the compared LR representations display close resolution values ($H[s]$) but significantly different numbers of clusters (K).

Figure 4.5, the distributions of centroids in the LR representations are compared with the distributions of frames in the HR representation; in each panel, points are coloured according to the value of the first diffusion coordinate in the LR representation. Visual inspection shows that the centroids obtained with average linkage are distributed in the LR diffusion-coordinate space in a way that is consistent with the distribution of HR frames. In both spaces, a clear colour gradient is visible along the x axis, indicating that neighbouring frames in the HR space are grouped into neighbouring clusters in the LR representation. For single linkage, by contrast, the relevance-maximising LR representation yields a point distribution that differs more noticeably from that of the HR frames. Moreover, the colouring suggests that frames that are distant in diffusion space are assigned to the same cluster.

This difference becomes even clearer when directly correlating the values of the diffusion coordinates in the HR and LR representations, as shown in the bottom-left panel of Figure 4.5. For both average and single linkage, a strong correlation is observed between the first diffusion coordinates in LR and HR, with Pearson coefficients of 0.95 and 0.85, respectively. Nevertheless, in the case of single linkage, some clusters contain frames spanning a broad range of HR diffusion-coordinate values, *i.e.*, frames carrying substantially different information are incorrectly grouped into the same bin. As expected, the random partition shows a complete lack of correlation with the reference HR diffusion coordinates, both in terms of point distribution and cluster composition.

4.4 Conclusions

This chapter examined the configurational Coarse-Graining (CG) of protein Molecular Dynamics (MD) trajectories from two complementary perspectives. First, it compared different agglomerative clustering procedures across a dataset of structurally distinct proteins, to assess how the resulting low-resolution representations depend on the linkage criterion and on the structural resolution used to define configurational similarity. Second, it analysed adenylate kinase in detail, to verify whether the clustering procedures favoured by the Resolution-Relevance (RR) analysis also preserve the structural organisation of the reference high-resolution configurational space. Corresponding results for the humanised IgG4 antibody are reported in Appendix C.

The first main result is that the different linkage criteria do not perform equally in the configurational CG problem. Across the 12-protein dataset, average, complete, weighted, and Ward linkage systematically yield positive relative Multiscale Relevance (MSR) values over different proteins and structural representations, whereas centroid and median linkage depend more strongly on the specific system and deteriorate when the description is reduced from all-atom to residue-level representations. This indicates that some clustering methods provide robust informative representations across scales, while others are less reliable as general CG

tools. Importantly, this distinction is not merely empirical. The best-performing methods largely coincide with those that induce a monotonic hierarchy during agglomeration and therefore an ultrametric structure in cluster space. Their success is thus consistent with the idea that protein configurational space is itself organised in a hierarchical, approximately self-similar way, so that clustering procedures able to preserve such an organisation are also those that yield the most informative low-resolution descriptions.

A separate discussion is required for single linkage. In the all-atom representation, the broader spread of RMSD values still allows this method to generate differentiated cluster populations. However, when the configurational description is reduced to C_α or C_β atoms, single linkage is strongly affected by the chaining effect and tends to merge frames that are not genuinely similar. As a consequence, its MSR becomes comparable to, or even lower than, the random baseline. This shows that even a formally hierarchical procedure can become uninformative when its definition of inter-cluster distance is not compatible with the structural resolution adopted. In this sense, single linkage represents an important exception: although it also induces an ultrametric structure, its poor performance indicates that hierarchy alone is not sufficient when the clustering rule is operationally dominated by the chaining effect.

The detailed analysis of adenylate kinase makes this picture more concrete. At fixed resolution, average linkage produces representations that are both simpler and more informative than those obtained with single linkage, reaching higher relevance with a smaller number of clusters. This ranking is confirmed by the structural diagnostics. In the average-linkage partitions, the inter-cluster covariance becomes larger than the intra-cluster contribution already at relatively small numbers of states, indicating that the retained representation captures more structural information than what remains within the discarded fluctuations. The diffusion-map analysis leads to the same conclusion, showing that the corresponding low-resolution manifolds preserve a substantially stronger correspondence with the high-resolution reference than those obtained through single or random clustering.

These results also clarify the role of RR framework in this problem. Relevance alone cannot be regarded as an absolute measure of the quality of a partition, because it depends only on cluster occupancies and is therefore blind to the geometric content of the grouping. Its usefulness emerges when the comparison is restricted to a physically meaningful family of partitions, in which case larger MSR values are associated with clustering procedures that better preserve the structural organisation of the reference configurational space. In this sense, the RR framework does not replace geometric analysis, but provides an unsupervised criterion to compare alternative configurational CG procedures.

More generally, the results of this chapter suggest that the same framework can be extended beyond the specific clustering protocols considered here. Since the RR analysis is able to discriminate between more and less informative configurational

partitions, it can also be used to compare different definitions of configurational similarity. This is particularly relevant because, although RMSD is the most common choice, its usefulness depends sensitively on the structural alignment procedure and it can be strongly influenced by flexible regions, so that alternative metrics may offer a more faithful description of conformational variability when the biologically relevant degrees of freedom are more specific. For example, in RNA systems, measures such as eRMSD, which focus on the relative arrangement of nucleobases rather than on a global Cartesian superposition, have been introduced precisely to better capture differences in base-pairing and stacking interaction patterns, and are therefore often more informative than standard RMSD when the structural question concerns RNA motifs, folding states, or interaction networks. Likewise, for DNA, descriptors based on helicoidal or backbone parameters can be more appropriate when the problem is to characterize sequence-dependent shape variations, groove width, or BI/BII backbone substates, rather than overall structural deviation.

This point also provides a natural connection to the next chapter. While the present analysis focused on configurational CG, *i.e.*, on how molecular configurations are grouped into states, the same general strategy can be extended to structural CG, where the question is no longer how to partition the sampled trajectory, but how much structural detail must be retained in order to preserve the informative properties of the molecule. In this sense, the next chapter shifts the focus from the CG of configurational space to the CG of the molecular representation itself.

Take-home message

Protein Molecular Dynamics trajectories can be compressed by clustering sampled conformations into a smaller number of configurational states. In this chapter, the informativeness of the resulting low-resolution representations was shown to depend strongly on the clustering protocol. By comparing different agglomerative clustering strategies, the analysis showed that Multi-scale Relevance identified the protocols that better preserved the geometry of the original configurational space. Informative configurational Coarse-Graining representations are therefore not arbitrary reductions of the trajectory, but partitions that retain the hierarchical organisation of the protein conformational landscape.

5

Structural Resolution

5.1 Introduction

In the previous chapter, the problem of **CG** was addressed from the standpoint of configurations, asking ourselves how the structures sampled along a **MD** trajectory could be grouped into informative states. The same perspective can be extended one step further by considering structural **CG**, in which the object to coarsen is no longer the configurational space, but rather the system structure itself, and the question is then how much structural detail must be retained for the low-resolution representation to remain informative.

Proteins are paradigmatic systems for which this question is particularly non-trivial. Their biological function is generally associated not only with the existence of a well-defined native structure, but also with a significant degree of flexibility, which allows them to explore a heterogeneous configurational space through motions spanning different spatial and temporal scales (Frauenfelder et al., 1991; Dill et al., 1997; Henzler-Wildman et al., 2007; Boehr et al., 2009; Dill et al., 2008). **MD** simulations represent an essential tool to investigate this interplay between structure and dynamics, since they provide an atomistic description of molecular motion (McCammon et al., 1977; Karplus et al., 2002; Dror et al., 2012). This level of detail is invaluable when a faithful microscopic characterisation of the system is required. For the purpose of gathering mechanistic insight in the system's function, however, a highly detailed representation is not necessarily a more informative one.

Indeed, an overly rich description may obscure the large-scale structural patterns and collective degrees of freedom most relevant to interpreting the sampled ensemble. As an extreme example, the precise knowledge of every atomic coordi-

This chapter is based on material originally published in M. Mele, R. Fiorentini, T. Tarenzi, G. Mattiotti and R. Potestio, *Determining the optimal structural resolution of proteins through an information-theoretic analysis of their conformational ensemble*, Journal of Chemical Theory and Computation, 22(3), 1244-1257 (2026). The material has been revised and adapted for consistency with the structure and notation of this thesis, and is reproduced in accordance with the terms of the applicable licence. The published version is available at: <https://doi.org/10.1021/acs.jctc.5c01773>.

nate is not necessarily required to distinguish between major conformational basins, since this task can often be accomplished through a much smaller number of suitably chosen coarse variables (Best et al., 2005; Peters et al., 2006; Noé et al., 2008; Chodera et al., 2014; Noé et al., 2017). This observation reflects a more general distinction between data and information: an atomistic trajectory contains a large amount of structural data, yet the challenge is not to preserve every microscopic fluctuation, but to identify a representation in which the relevant organisation of configurational space emerges clearly from the ensemble.

This issue is central in CG modelling. A reduced representation should be simple enough to suppress irrelevant microscopic fluctuations, yet detailed enough to retain the structural and dynamical features that characterise the molecule (Noid, 2013; Saunders et al., 2013; Kmiecik et al., 2016). In practice, however, the choice of structural resolution is often guided by empirical intuition, computational convenience, or the conventions of established models. Widely used protein CG force fields, such as Martini and SIRAH, provide successful examples of intermediate-resolution representations capable of reproducing many relevant conformational and thermodynamic properties while substantially reducing computational cost (Souza et al., 2021; Machado et al., 2019; Klein et al., 2023). Yet, their effectiveness does not by itself answer a more general question: whether it is possible to determine, directly from the data, the structural resolution at which a protein admits its most parsimonious informative description.

The construction of a structural CG representation involves two conceptually distinct questions. The first concerns the *resolution* of the representation, namely, how many structural degrees of freedom should be retained. The second concerns the *mapping*, that is, which specific atoms should be kept to realise that representation. These two aspects are often intertwined in practical applications, but they need not coincide conceptually. In particular, before asking which mapping is most appropriate, one may ask whether the data themselves contain an intrinsic indication of the amount of structural detail that must be preserved. This distinction is especially relevant in light of previous work on mapping entropy and mapping optimisation, where the focus is on identifying which degrees of freedom are most informative once a target resolution has been fixed (Shell, 2008; Chaimovich et al., 2011; Giulini et al., 2020; Giulini et al., 2021; Holtzman et al., 2022; Giulini et al., 2024).

The present chapter addresses precisely the first of these two questions. Building on the RR framework introduced in Section 2.3, we consider the problem of identifying the structural resolution at which a protein conformational ensemble admits its most parsimonious yet informative description. The key idea is that a low-resolution structural representation induces a corresponding clustering of the sampled configurations, and that the statistical properties of this clustering depend on the number of atoms retained in the mapping. By scanning different values of the retained-atom number and quantifying the resulting balance between re-

tained information and compression level, it is possible to identify an optimal resolution scale in a fully data-driven, unsupervised way. This is the same general logic that, in the previous chapters, was used to characterise the size of informative low-resolution representations of the configurational space; here, it is transferred to the level of the structural description itself.

Within this framework, the chapter has two main objectives. The first is to assess whether the optimal structural resolution shows systematic trends across proteins of different sizes and dynamical characters. The second is to clarify which physical information is captured by the identified optimal scale by comparing it with independent structural diagnostics, such as the covariance of atomic fluctuations and the mapping entropy. As shown, the optimal number of retained atoms is not determined solely by size, but also by the extent and type of conformations sampled by the system.

5.2 Materials and Methods

5.2.1 Systems and Simulation Data

The analysis was carried out on a dataset of 11 proteins differing in size, structure, and function, together with the enzyme adenylate kinase (ADK), which was treated separately to investigate how optimal structural resolution depends on the conformational state explored. The 11-protein dataset was constructed following the pipeline introduced in [Tarenzi et al., 2022](#). Specifically, a representative, heterogeneous subset of proteins was selected from a larger database based on structural and dynamical criteria. The proteins considered here have PDB codes 1DSL, 1NOA, 1SNO, 1UNE, 1XWL, 1IGD, 1HYP, 1KNT, 1QKE, 2EXO, and 1KOE.

In addition to this benchmark set, adenylate kinase was used as a test case to probe the role of conformational variability explicitly. ADK is particularly suited for this purpose because of its pronounced large-scale structural rearrangements, which can be sampled on time scales accessible to standard atomistic MD simulations. The starting structure for the simulations corresponds to the crystallographic open form (PDB ID: 4AKE).

All simulations were performed with Gromacs 2018 ([Abraham et al., 2015](#)). The protein topology was described with the AMBER99SB-ILDN force field ([Lindorff-Larsen et al., 2010](#)), while water molecules were modelled with the TIP3P model ([Jorgensen et al., 1983](#)). Sodium and chloride ions were added to neutralise the system's total charge and to reproduce a physiological salt concentration of 150 mM. After energy minimisation, NVT and NPT equilibration stages were performed using the velocity-rescale thermostat ([Bussi et al., 2007](#)) and the Parrinello–Rahman barostat ([Parrinello et al., 1981](#)), respectively. A cut-off of 1.2 nm was used for van der Waals interactions and for the short-range component of the Coulomb interaction, whereas long-range electrostatics were treated with the Particle Mesh Ewald

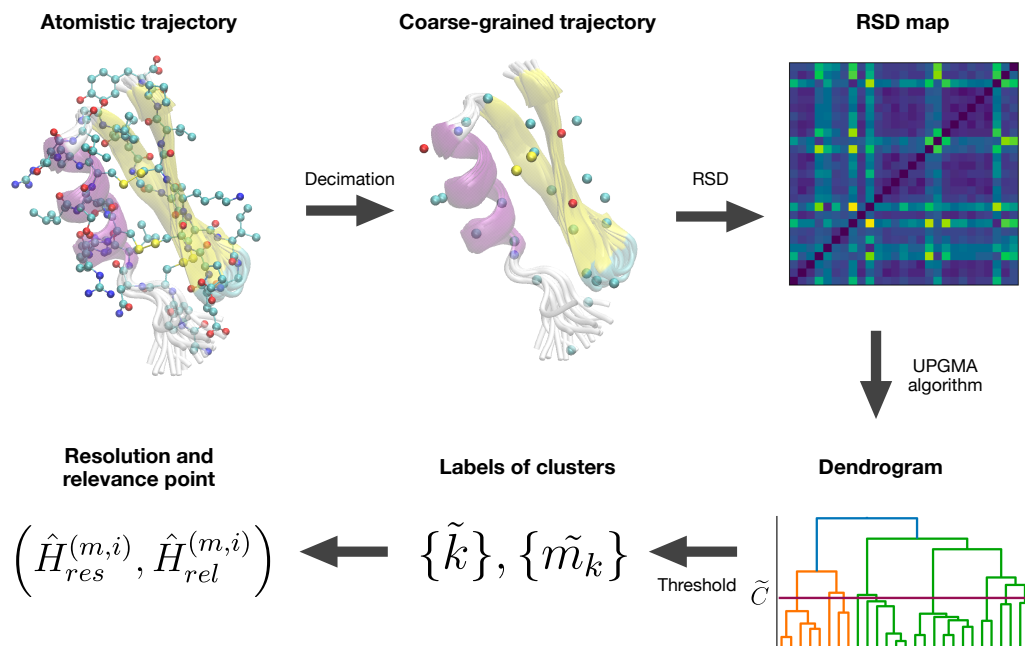


Figure 5.1: Infographic that illustrates the steps of the first part of the PROPRE protocol: the construction of the *RR* scatter plot. After filtering the atomistic trajectory (by choosing a random decimation mapping of the heavy atoms), the RSD map is built in order to perform a clustering of the filtered configurations. Each mapping corresponds to a point in the *RR* plane; by varying the selection of retained atoms and the number of retained atoms, one can reconstruct the curve as a scatter plot.

method (Darden et al., 1993; Essmann et al., 1995). Bond constraints involving hydrogen atoms were enforced through the LINCS algorithm (Hess et al., 1997), and the equations of motion were integrated with a time step of 2 fs. For each protein in the data set, a 500 ns trajectory was generated, and the first 100 ns were discarded as equilibration.

5.2.2 Structural Decimation and Induced Configurational Clustering

The analysis starts from the atomistic trajectory of a protein, described solely by the coordinates of its heavy atoms. If n_{ha} is the total number of heavy atoms, each configuration is represented as a point in a $3n_{\text{ha}}$ -dimensional space. Low-resolution structural representations were generated through decimation mappings, obtained by selecting a subset of N_{CG} heavy atoms out of the n_{ha} available ones. For a fixed value of N_{CG} , different mappings correspond to different subsets of retained atoms and therefore to different projected representations of the same trajectory. The overall workflow, from structural decimation to the construction of the induced configurational partition and of the corresponding *RR* point, is schematically shown in Figure 5.1.

For each mapping, the atomistic trajectory was projected onto the coordinates of the selected atoms and clustered by agglomerative hierarchical clustering using the

UPGMA algorithm (Sokal et al., 1958). The distance between two configurations was measured through the root squared distance (RSD),

$$\text{RSD}(\mathbf{X}^a, \mathbf{X}^b) = \left(\sum_{i=1}^n |\mathbf{x}_i^a - \mathbf{x}_i^b|^2 \right)^{1/2}, \quad (5.1)$$

which is proportional to the usual root mean square deviation, $\text{RSD} = \sqrt{n} \text{RMSD}$. The use of RSD is convenient here because, unlike RMSD, it preserves the absolute scale of the distance when the number of retained atoms changes, making it possible to adopt the same clustering threshold at all values of N_{CG} .

The clustering threshold, $\tilde{C}$, was defined from the all-atom trajectory as the largest value below which all sampled configurations remain distinguishable, so that each frame still forms a singleton cluster. The same threshold was then used to cut the dendrograms obtained from the projected trajectories at all values of N_{CG} . In this way, two reduced configurations were regarded as indistinguishable only if their retained structural differences were smaller than the minimum distance required to resolve distinct configurations in the full heavy-atom space.

For each pair given by a retained-atom number N_{CG} and a specific decimation mapping, this procedure yields a partition of the trajectory into clusters of structurally equivalent reduced configurations and therefore a point in the **RR** plane. The comparison of these points across different mappings and N_{CG} values provides the basis for identifying the optimal structural resolution discussed in the next subsection.

5.2.3 PROPRE Analysis and Identification of the Optimal Resolution

PROPRE, short for PROtein PReservEd sites, denotes the atom-selection procedure used to identify the most informative retained sites at each resolution. For each system, the PROPRE analysis was performed on 1000 frames extracted from the trajectory. Two sampling strategies were considered: uniform sampling, obtained by selecting equally spaced frames along the trajectory, and centroid sampling, obtained from the centroids of 1000 UPGMA clusters. For a fixed number of retained atoms N_{CG} , 50 random decimation mappings were generated, and the procedure described above was applied to each of them. The number of retained atoms was then progressively decreased with a step corresponding to 0.5% of the total number of heavy atoms of the protein. In this way, each system is associated with a family of **RR** points, obtained by scanning both the structural resolution and the specific choice of retained atoms.

The resulting **RR** curve was then smoothed to identify the characteristic resolutions; the exact protocol is described in Appendix A. Two reference points were considered: the value of N_{CG} at which the relevance reaches its maximum, denoted here by $N_{\text{opt}}^{\text{MR}}$, and the value at which the local slope of the curve is equal to -1 , denoted by $N_{\text{opt}}^{\text{IT}}$.

For adenylate kinase, the same procedure was applied not only to the full trajectory, but also separately to the portions associated with the open and closed conformational states, identified through UPGMA clustering. This enabled comparison of optimal structural resolution across different regions of the conformational landscape.

5.2.4 Structural Fluctuation Analysis

To quantify the structural fluctuations of a molecular system composed of n_{ha} heavy atoms over T time steps, we compute the covariance matrix of atomic positions. Let $\mathbf{P} \in \mathbb{R}^{T \times 3n_{\text{ha}}}$ denote the matrix of atomic configurations, where each row corresponds to the 3D positions of all heavy atoms at a given time step, flattened into a vector of length $3n_{\text{ha}}$. The mean configuration over time is given by $\bar{\mathbf{P}} \in \mathbb{R}^{1 \times 3n_{\text{ha}}}$. The covariance matrix $\mathbf{C} \in \mathbb{R}^{3n_{\text{ha}} \times 3n_{\text{ha}}}$ is then defined as:

$$\mathbf{C} = \frac{1}{T-1} (\mathbf{P} - \bar{\mathbf{P}})^{\top} (\mathbf{P} - \bar{\mathbf{P}}). \quad (5.2)$$

This matrix captures the variance and covariance of positional fluctuations, including both intra- and inter-atomic correlations among Cartesian coordinates. As a scalar descriptor of the overall fluctuation magnitude, we consider the *trace* of the covariance matrix, defined as the sum of its diagonal elements:

$$\mathcal{T}^{\text{ha}} = \text{Tr}(\mathbf{C}) = \frac{1}{T-1} \sum_{i=1}^T \|\mathbf{P}_i - \bar{\mathbf{P}}\|^2. \quad (5.3)$$

The same construction can be applied when only N_{CG} atoms are kept, chosen at random among the n_{ha} heavy atoms. The corresponding fluctuation measure is denoted $\mathcal{T}^{\text{CG}}$. Since the trace grows proportionally with the number of atoms, the two quantities $\mathcal{T}^{\text{ha}}$ and $\mathcal{T}^{\text{CG}}$ are not directly comparable. We therefore introduce a size-independent observable

$$\text{cov} = \frac{\mathcal{T}^{\text{CG}} \cdot n_{\text{ha}}}{\mathcal{T}^{\text{ha}} \cdot N_{\text{CG}}}. \quad (5.4)$$

This normalisation accounts for the difference in system size by assigning an average fluctuation contribution to the atoms excluded during CG. By construction, $\text{cov} = 1$ when all atoms are retained, and values near 1 indicate that the retained subset captures the overall structural variability of the whole system. The behaviour of cov as a function of N_{CG} provides a diagnostic for quantifying how representative an atom's subset is of the system's fluctuations.

5.2.5 Mapping Entropy

Mapping entropy quantifies the intrinsic loss of information that arises when a fine-grained system is projected onto a CG representation. This concept was first intro-

duced within the framework of the relative entropy method (Shell, 2008) and has been widely applied to the analysis of proteins and other systems (Giulini et al., 2020; Holtzman et al., 2022; Giulini et al., 2021; Aldrigo et al., 2025).

Let $\mathbf{x} \in \mathbb{R}^{3n_{\text{ha}}}$ denote a microscopic configuration of a molecular system composed of n_{ha} heavy atoms, sampled from the canonical distribution $p(\mathbf{x}) \propto e^{-\beta U(\mathbf{x})}$, where $U(\mathbf{x})$ is the potential energy and $\beta = 1/(k_B T)$ is the inverse temperature. A CG representation can be defined by selecting a subset of $N_{\text{CG}} < n_{\text{ha}}$ atoms. The selection is encoded by a binary vector $\sigma = \{\sigma_i \in \{0, 1\}\}_{i=1}^{n_{\text{ha}}}$ such that $\sum_{i=1}^{n_{\text{ha}}} \sigma_i = N_{\text{CG}}$. This defines a projection operator M that maps the full configuration $\mathbf{x}$ to a CG macrostate $\mathbf{X} = M(\mathbf{x})$, retaining only the positions of the selected atoms.

The mapping entropy S_{map} is defined as the Kullback–Leibler divergence between the original microscopic distribution $p(\mathbf{x})$ and the reconstructed distribution $\tilde{p}(\mathbf{x})$ inferred from the CG representation:

$$S_{\text{map}} = k_B \int p(\mathbf{x}) \ln \left(\frac{p(\mathbf{x})}{\tilde{p}(\mathbf{x})} \right) d\mathbf{x}. \quad (5.5)$$

The reconstructed distribution $\tilde{p}(\mathbf{x})$ corresponds to the statistical model obtained by backmapping from the CG variables and is given by

$$\tilde{p}(\mathbf{x}) = \frac{p_{\text{CG}}(M(\mathbf{x}))}{\Omega(M(\mathbf{x}))}, \quad (5.6)$$

where $p_{\text{CG}}(\mathbf{X}) = \int p(\mathbf{x}') \delta(M(\mathbf{x}') - \mathbf{X}) d\mathbf{x}'$ is the induced probability over CG configurations, and $\Omega(\mathbf{X}) = \int \delta(M(\mathbf{x}') - \mathbf{X}) d\mathbf{x}'$ is the degeneracy of the macrostate $\mathbf{X}$, i.e., the volume of microscopic configurations such that $M(\mathbf{x}) = \mathbf{X}$.

This functional is non-negative and vanishes only when the CG representation preserves the full statistical structure of the original representation. In this work, we evaluate S_{map} for a sample of randomly generated selections of N_{CG} atoms, drawn from the n_{ha} available atoms, to estimate the average information loss associated with unsupervised, non-optimised CG choices.

Mapping entropy calculations were performed using the EXCOGITO software package (Giulini et al., 2024), a general-purpose toolkit for CG analysis. In our case, the mapping entropy was estimated as the Kullback–Leibler divergence between the full microscopic distribution and the reconstructed one, with structural similarity assessed through clustering based on the root square deviation (RSD), adopting the same distance threshold as in the PROPRE protocol. Further details on the methodology can be found in (Giulini et al., 2024; Holtzman et al., 2022; Shell, 2008; Chaimovich et al., 2011).

5.3 Results

The results presented in this section address three related aspects of the structural CG problem. First, the 11-protein dataset is used to determine whether the opti-

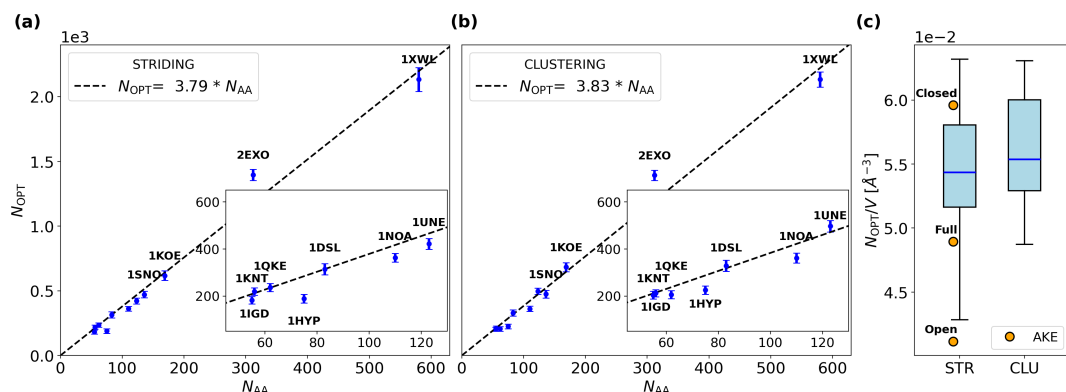


Figure 5.2: (a, b) Scatter plots illustrating the number of optimal sites identified by PROPRE for each protein in relation to the number of residues; for both cases (striding and clustering), the correlation coefficient of the linear fit (R^2) is 0.98, and the Pearson correlation coefficient is 0.99. (a) Analysis based on 1000 equidistant frames extracted through striding. (b) Analysis based on the centroids of 1000 clusters obtained from a UPGMA clustering procedure. Error bars indicate the standard deviation in the number of optimal sites within the most probable bin, as determined by PROPRE using the density-based protocol. (c) Box plots of the ratio between the number of optimal sites and the average sampled volume (N_{OPT}/V) for the two strategies: striding (STR) and clustering (CLU). For STR, orange points indicate values for the AKE protein computed from the full trajectory, as well as separately for the open and closed conformations.

mal number of retained atoms displays a systematic dependence on protein size. Second, adenylate kinase is employed as a case study to assess how the optimal structural resolution depends on the explored conformational state. Third, the **RR**-based optimality estimates are compared with two independent structural diagnostics, namely the covariance of atomic fluctuations and the mapping entropy, to clarify which features of the conformational ensemble are captured by the identified resolution scale.

5.3.1 Optimal Number of Sites *vs.* Number of Protein Residues

The PROPRE algorithm allows the identification of the optimal structural resolution of a specific protein, where the optimal resolution is defined as the number of heavy atoms that must be retained to minimise information loss upon **CG**. In the following, the optimal number of retained atoms is primarily identified with the point of slope -1 along the **RR** curve, namely $N_{\text{OPT}}^{\text{IT}}$. The complementary estimates obtained from the maximum of the relevance, $N_{\text{OPT}}^{\text{RM}}$, together with the extended fit on the 11-protein dataset and the quantitative comparison with Martini and SIRAH, are reported in **Appendix D**.

It is natural to expect that the optimal number of retained atoms depends on the specific system under investigation and on its features, such as size, structure, and flexibility. To assess whether a more general trend is present, the PROPRE analysis was first applied to the trajectories of the 11 proteins of the dataset introduced in **Section 5.2.1**. The resulting relation between the optimal number of retained atoms and the total number of amino acids is reported in **Figure 5.2a,b** for the two

trajectory-sampling strategies considered here, namely striding and centroid selection through UPGMA clustering.

Despite the large differences in structure, size, and dynamical features across the dataset, a clear linear trend emerges in both cases. The same analysis extended to the relevance-maximising point, $N_{\text{OPT}}^{\text{RM}}$, is reported in [Appendix D](#), where it is shown that the linear scaling with protein size is preserved also under this alternative operational definition of the optimal resolution. Larger proteins require a proportionally larger number of retained atoms to preserve an informative representation of their configurational space. The existence of this correlation is not in itself surprising, since larger systems are expected to require more structural detail. However, the fact that the dependence is so clearly linear across a set of structurally and dynamically heterogeneous proteins is far from trivial. In both sampling protocols, the fitted slope is close to four retained heavy atoms per residue, indicating that the optimal representation is substantially more detailed than a one-site-per-residue description, while remaining much more compact than the full atomistic one.

This result suggests that, for globular proteins that explore relatively limited fluctuations around their native structures, the minimally informative representation corresponds approximately to a backbone-level description supplemented by limited side-chain detail. In this sense, the analysis identifies a characteristic structural scale at which a large fraction of microscopic detail can be removed without suppressing the informative organisation of the conformational ensemble.

The consistency between the two sampling strategies in [Figure 5.2a,b](#) is also noteworthy. In the present dataset, none of the proteins undergoes major conformational rearrangements, which likely explains why uniform and centroid sampling yield nearly identical optimal resolutions. Differences between the two procedures are expected to become more pronounced when analysing trajectories characterised by greater heterogeneity in configurational states.

A complementary way of looking at the same trend is reported in [Figure 5.2c](#), where the optimal number of retained atoms is normalised by the average configurational volume explored during the simulation. The resulting values of N_{OPT}/V are narrowly distributed across the 11 proteins, despite their differences in size and topology. This suggests the presence of a characteristic density, indicating that proteins with stable globular folds tend to retain a similar number of optimal sites per unit of sampled volume.

To explicitly probe the role of conformational variability, adenylate kinase was considered as a separate case study. The same protein was analysed in three different ways: using the full trajectory, using only the subset corresponding to the open state, and using only the subset corresponding to the closed state, the latter two obtained through UPGMA clustering as described in [Section 5.2.3](#). As shown by the orange points in [Figure 5.2c](#), all three adenylate kinase cases fall within the same density range as the 11 proteins, indicating that variations in the sampled vol-

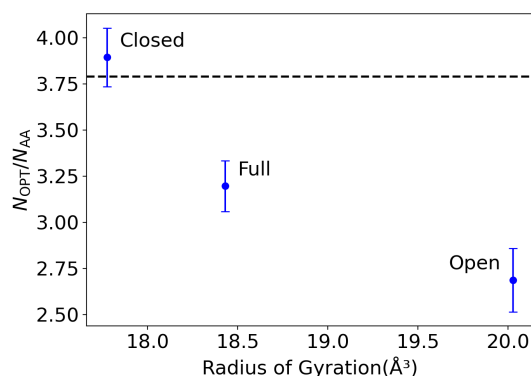


Figure 5.3: Scatter plot of the ratio between the optimal site number and residue number ($N_{\text{OPT}}/N_{\text{AA}}$) identified for adenylate kinase from the full trajectory and from the subsets corresponding to the open and closed states, as a function of the average radius of gyration sampled in each case. The optimal atom number corresponds to the resolution point of slope -1 . The dashed horizontal line indicates the reference value fitted on the 11-protein dataset.

ume are accompanied by corresponding changes in the optimal number of retained atoms.

A complementary insight is obtained by considering the ratio between the optimal number of retained atoms and the number of amino acids as a function of the average radius of gyration, reported in Figure 5.3. Only the closed-state trajectory aligns with the reference value fitted on the 11-protein dataset, indicated by the dashed horizontal line. By contrast, the full trajectory and the open-state subset yield systematically lower values of $N_{\text{OPT}}^{\text{IT}}/N_{\text{AA}}$.

These results indicate that the optimal number of retained atoms decreases as the average radius of gyration increases. In other words, conformational ensembles spanning broader configurational regions require fewer atoms to be retained, whereas subsets confined to smaller fluctuations need more atoms to achieve the same information balance. This inverse relation is consistent with the idea that large-scale, highly collective motions can be described with a reduced number of degrees of freedom, while small-amplitude local fluctuations require a higher structural resolution to be meaningfully distinguished.

In the specific case of adenylate kinase, the open-to-closed transition is known to be dominated by hinge-bending motions of the LID and NMP domains, that is, by a collective and intrinsically low-dimensional rearrangement. The present results quantitatively support this picture: the broader the conformational exploration, the fewer the atomic degrees of freedom needed for an informative representation. Conversely, when the analysis is restricted to a smaller conformational basin, the remaining variability is dominated by finer local rearrangements, and a higher structural resolution is required to preserve structural diversity.

Overall, the results discussed so far show that system size is a key determinant of the optimal structural resolution, but it is not the only one. The optimal number of retained atoms also depends on intrinsic dynamical features, such as the extent of conformational variability and the diversity of the structural states explored by the system.

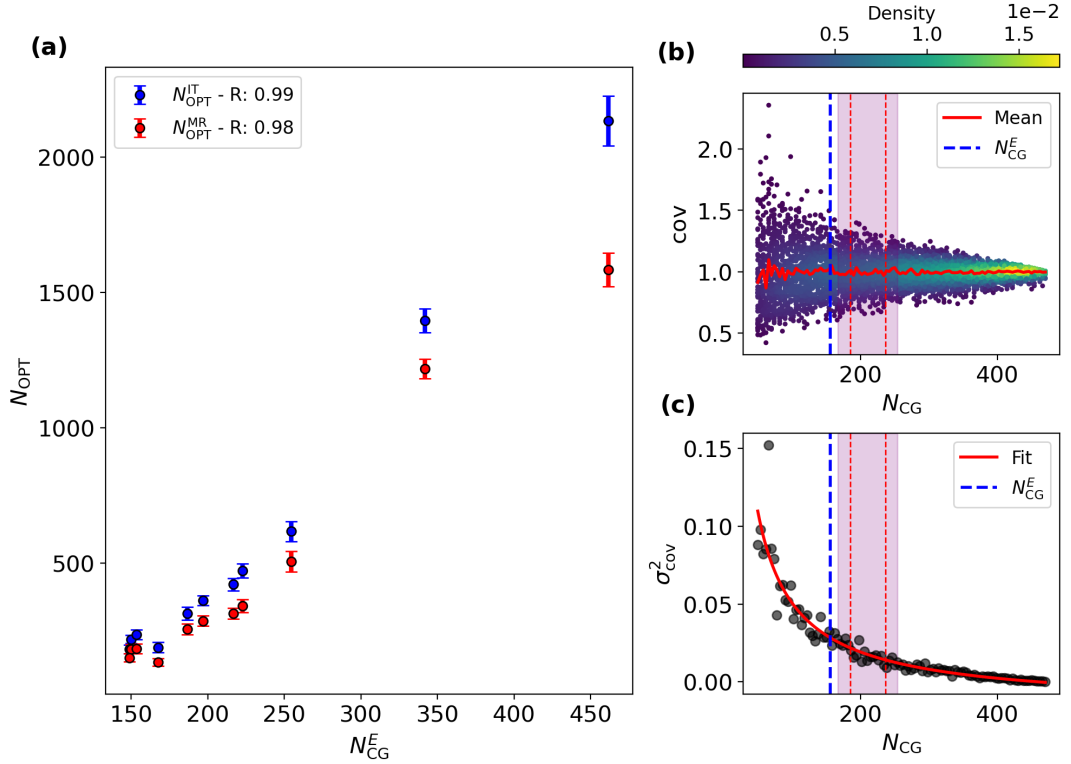


Figure 5.4: (a) Correlation between the elbow position (N_{CG}^E) in the variance of the CG covariance (σ_{cov}^2) and the optimal number of retained atoms N_{OPT} identified by PROPRES, across a dataset of 11 proteins. Blue markers correspond to the information-theoretic estimate N_{OPT}^{IT} (slope $\mu = -1$), red markers to the relevance-maximum estimate N_{OPT}^{MR} (slope $\mu = 0$). The legend reports the Pearson correlation coefficient for each criterion, and error bars indicate the standard deviation of N_{OPT} within the most-probable bin returned by PROPRES. (b-c) Analysis of protein 1IGD. In both panels, the vertical blue dashed line marks the elbow position (N_{CG}^E), the vertical dashed red lines delimit the core PROPRES interval, and the shaded region extends it by one standard deviation. (b) Distribution of the normalised CG covariance cov over 50 random atom selections for each N_{CG} ; colour encodes the pointwise density, the solid red line traces the running average. (c) Empirical variance $\sigma_{cov}^2(N_{CG})$ (grey dots) and fitted power-law decay (red curve).

5.3.2 Optimal Number of Sites vs. Position Covariance

To further investigate the factors influencing optimal resolution, its relationship with the distribution of atomic displacements was examined through the normalised trace of the covariance matrix, cov , introduced in Section 5.2.4. For each value of N_{CG} , cov was computed over 50 randomly selected subsets of heavy atoms.

The results are summarised in Figure 5.4. The right panels refer to the representative protein 1IGD and illustrate the behaviour of the observables and the procedure used to extract the characteristic structural resolution. The upper panel, Figure 5.4b, shows that cov fluctuates around 1 over the whole range of N_{CG} , indicating that, on average, the overall magnitude of the structural fluctuations is preserved even in reduced representations. At the same time, the spread of the values around this average decreases systematically with increasing N_{CG} , reflecting the reduced variability among different atom selections as the representation becomes

sufficiently detailed to capture most of the fluctuation profile.

To quantify this trend, the variance $\sigma_{\text{cov}}^2(N_{\text{CG}})$ was computed for each value of N_{CG} . As shown in Figure 5.4c, this quantity displays an elbow-shaped decay, suggesting that beyond a certain structural resolution, the addition of further atoms contributes only marginally to reproducing the fluctuations of the full system. The elbow position, denoted as N_{CG}^E , identifies the point at which the CG description becomes sufficiently rich to capture the dominant configurational variability.

The shaded region in Figure 5.4c marks the optimal interval identified by PROPRE, delimited by $N_{\text{OPT}}^{\text{RM}}$ and $N_{\text{OPT}}^{\text{IT}}$. Notably, this interval lies systematically just beyond the elbow. This indicates that PROPRE tends to select resolutions slightly higher than those strictly required to retain the dominant structural fluctuations. In other words, the optimal resolution is not simply the minimum needed to preserve fluctuation amplitudes, but rather a slightly finer representation at which the reduced structure becomes both fluctuation-preserving and statistically informative.

This trend extends to the whole dataset, as shown in Figure 5.4a, where the elbow positions extracted from the covariance analysis are compared with the optimal resolutions identified by PROPRE. A strong linear correlation is observed for both definitions of the optimum, with Pearson coefficients larger than 0.98. These results show that the optimal resolution identified through the RR analysis systematically lies just beyond the point at which the covariance-based description begins to saturate.

5.3.3 Optimal Number of Sites *vs.* Mapping Entropy

A complementary analysis was carried out by comparing the PROPRE estimates with the mapping entropy S_{map} , introduced in Section 5.2.5, which quantifies the information loss associated with the projection of the full system onto a reduced structural representation. In the present case, S_{map} was computed for randomly selected subsets of retained atoms, to characterise the information loss associated with non-optimised CG mappings. For each value of N_{CG} , 50 random subsets of heavy atoms were considered.

For the representative protein 1IGD, the distribution of S_{map} is reported in Figure 5.5b. Its average value decreases smoothly as N_{CG} increases, indicating that, on average, the reconstruction quality improves progressively as more atoms are retained. However, this mean behaviour does not exhibit a distinct elbow or plateau and therefore does not by itself identify a special resolution threshold.

A more informative quantity is instead the variance of the mapping entropy across different random selections. As shown in Figure 5.5c, $\sigma_{\text{Smap}}^2(N_{\text{CG}})$ exhibits a well-defined maximum at an intermediate value of N_{CG} , denoted as N_{CG}^P . This behaviour has a clear interpretation. At very low values of N_{CG} , almost all mappings are poor and therefore similarly uninformative. At very high values of N_{CG} , almost all mappings are good and therefore similarly accurate. Between these two limits,

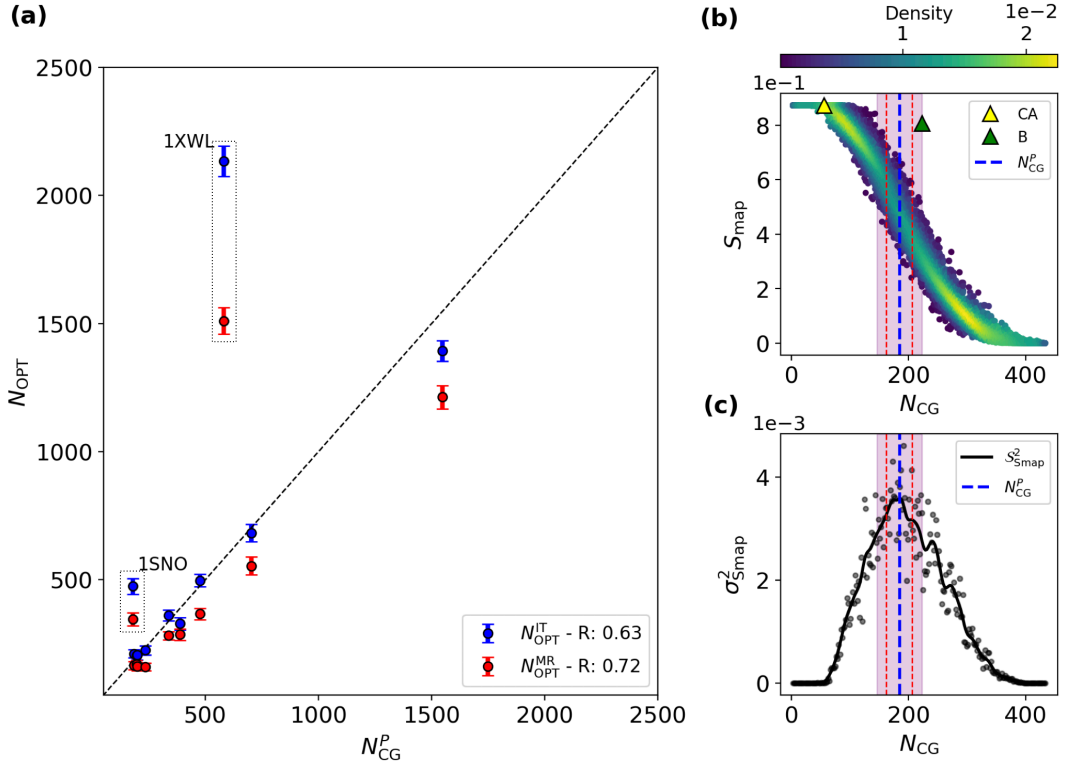


Figure 5.5: (a) Correlation between the peak position in the variance of the mapping entropy (N_{CG}^P) and the optimal number of retained atoms N_{OPT} identified by PROPRES, across a dataset of 11 proteins. Blue markers correspond to the information-theoretic estimate N_{OPT}^{IT} (slope $\mu = -1$), red markers to the relevance-maximum estimate N_{OPT}^{MR} (slope $\mu = 0$). The legend reports the Pearson correlation coefficient for each criterion, and error bars indicate the standard deviation of N_{OPT} within the most-probable bin returned by PROPRES. (b-c) Analysis of protein 1IGD. In both panels, the vertical blue dashed line marks the peak position N_{CG}^P , vertical dashed red lines delimit the core PROPRES interval, and the shaded region extends it by one standard deviation. (b) Distribution of the mapping entropy S_{map} over 50 random atom selections for each N_{CG} ; color encodes the pointwise density. The yellow and green triangles highlight the values of S_{map} obtained from the specific selections of all C_α atoms and all backbone atoms, respectively. (c) Empirical variance $\sigma_{Smap}^2(N_{CG})$ (gray dots) and its Gaussian-smoothed estimate S_{Smap}^2 (black line).

the specific choice of retained atoms matters the most, and this produces a peak in the variance.

The relation between this characteristic point and the PROPRES optimum is reported in Figure 5.5a. Also in this case, a positive correlation emerges between the peak position in the variance of the mapping entropy and the optimal number of retained atoms identified by PROPRES. Thus, the optimal structural resolution falls in the regime where the reduced representation is most sensitive to which atoms are actually retained.

An additional indication comes from the specific chemically motivated selections also reported in Figure 5.5b. The values of S_{map} obtained by retaining all C_α atoms or all backbone atoms lie well above those corresponding to random heavy-atom subsets of the same size. This shows that, for the present dataset, these standard structural reductions are not especially informative. Since the proteins consid-

ered here do not undergo large conformational rearrangements but rather fluctuate within local basins around their native globular structures, the relevant structural information is not fully captured by C_α or backbone coordinates alone.

Taken together, the covariance-based and mapping-entropy-based analyses provide two complementary interpretations of the same result. On the one hand, the optimal interval identified by PROPRE lies just beyond the minimum resolution needed to retain the dominant structural fluctuations. On the other hand, it lies in a regime where the specific choice of retained atoms matters the most in terms of information loss. The structural resolution identified by PROPRE therefore emerges as a minimal yet informative representation of the conformational ensemble, determined jointly by system size and by the nature of the configurational exploration.

5.4 Conclusions

In this chapter, the **RR** framework has been extended from the analysis of configurational representations to the study of structural **CG** in proteins. The central question was no longer how many configurational states are required to describe a system, but rather how many atomic degrees of freedom must be retained to preserve an informative representation of its conformational ensemble. Within this perspective, the PROPRE procedure provides an operational criterion for directly identifying the optimal structural resolution of a protein from simulation data.

The results presented in this chapter show that the optimal number of retained atoms is not arbitrary but depends systematically on the system's properties and the type of conformational variability explored. For the 11 proteins considered here, the optimal structural resolution scales approximately linearly with the number of amino acids, indicating that larger systems require proportionally more retained atoms to preserve an informative description of their configurational space. At the same time, the analysis of adenylate kinase shows that this dependence is not solely determined by size but is also influenced by the conformational region sampled by the system. In particular, broader and more collective motions require fewer retained atoms, whereas more localised fluctuations within restricted basins require a finer structural description. This is consistent with the broader view that protein dynamics can often be described in terms of collective variables and reduced dynamical representations (Noé et al., 2017; Chodera et al., 2014; Noé et al., 2008).

A further relevant result is that the structural resolution identified by PROPRE is quantitatively close to those adopted by widely used **CG** models such as Martini and SIRAH (Souza et al., 2021; Klein et al., 2023; Machado et al., 2019). More specifically, for the proteins considered in this chapter, the optimal number of retained sites predicted by PROPRE falls systematically between the number of sites associated with Martini-like and SIRAH-like representations, as discussed quantitatively in [Appendix D](#). In this sense, the present analysis does not merely recover a generic intermediate level of **CG**, but identifies a structural resolution that is directly com-

parable with the one employed by two of the most widely used protein CG force fields.

This result is particularly significant because Martini and SIRAH represent two well-established and successful examples of intermediate-resolution protein CG modelling, which differ in both mapping strategy and level of structural detail (Souza et al., 2021; Klein et al., 2023; Machado et al., 2019). The fact that the optimal resolution identified here lies between them provides a non-trivial, data-driven indication that the structural compression selected by PROPRE corresponds to a physically meaningful regime, rather than to an arbitrary decimation threshold. In other words, the analysis places the optimal information-theoretic resolution in the same structural range that has proved effective in practical CG descriptions (Noid, 2013; Saunders et al., 2013; Kmiecik et al., 2016).

At the same time, this comparison should be interpreted at the resolution level, not at the mapping level. Martini and SIRAH employ chemically informed and non-uniform representations, so that their effective descriptive power is richer than what is captured by the nominal number of sites alone. By contrast, PROPRE identifies a global optimal structural scale, averaged over decimation mappings. The agreement should therefore not be read as a validation of a specific force field, but rather as an independent rationale for why protein representations at this intermediate resolution are especially effective for the class of systems considered here. The robustness of the identified optimal resolution with respect to stochastic mapping generation and independent simulation replicas is assessed in Appendix D, confirming that the observed variability remains within the intrinsic uncertainty of the procedure.

The comparison with the covariance of atomic fluctuations and with the mapping entropy further clarifies the meaning of the identified optimal scale. On the one hand, the PROPRE optimum lies just beyond the minimum structural resolution needed to reproduce the dominant positional fluctuations of the full system. On the other hand, it lies in the regime where the information loss associated with structural projection is most sensitive to the specific degrees of freedom that are retained. Taken together, these results indicate that the optimal structural resolution identified by the RR analysis corresponds to a minimal yet informative representation of the conformational ensemble. This interpretation is consistent with the information-theoretic picture underlying relative-entropy and mapping-entropy approaches to CG (Shell, 2008; Chaimovich et al., 2011; Holtzman et al., 2022; Giulini et al., 2020; Giulini et al., 2021).

More generally, the analysis developed in this chapter shows that the balance between compression and information retention can also be used to address the choice of structural resolution in molecular systems. In this sense, the RR framework is extended beyond the partitioning of configurational space and provides a principled criterion for identifying reduced structural representations directly from simulation data. The results obtained for the protein systems considered in this chapter

support the idea that the optimal resolution is an intrinsic property of the sampled conformational ensemble, rather than an arbitrary modelling choice (Marsili et al., 2022; Cubero et al., 2019; Holtzman et al., 2022).

Take-home message

A protein can be represented at different structural resolutions, depending on how many residues or interaction sites are retained in the reduced description. This chapter investigated whether an optimal number of retained sites could be identified by analysing the configurational partitions induced by structural decimation. The results showed that the optimal structural resolution scaled approximately linearly with protein size and correlated with both collective fluctuations and Mapping Entropy. This suggests that an informative protein representation should not be chosen only by chemical intuition or fixed conventions, but can be also selected from the statistical information retained by the reduced structural description.

Networks

6

Theoretical Background

Between 2023 and 2024, the most advanced **Artificial Intelligence (AI)** models exhibited substantial gains on major benchmark tasks, with score increases of 18.84 percentage points in multimodal reasoning, 48.9 points on expert-level scientific questions, and 67.3 points on software engineering tasks (Maslej et al., 2025). In parallel, **AI** applications have expanded far beyond controlled research settings. In healthcare, for example, the number of FDA-approved **AI**-enabled medical devices grew from 6 in 2015 to 223 in 2023, while autonomous-driving services moved from experimental testing at the beginning of 2024 to deployment at scales exceeding 150,000 rides per week only in the US by late 2024 (Maslej et al., 2025). The same trend is visible in scientific activity, where **AI**-related publications have more than doubled over the last decade, and the share of computer-science research devoted to **AI** is steadily increasing (Maslej et al., 2025). Taken together, these trends capture both the scale and the speed of what is often defined as the **Artificial Intelligence** revolution. Neural networks and related learning systems are currently reshaping a broad range of domains, from everyday technologies to industry, from education to scientific research, where they are increasingly used as standard tools for the analysis, generation, and organisation of complex data.

However, despite this extraordinary success, the rules governing the behaviour of **AI** models remain only partially understood (Goodfellow et al., 2016; Zdeborová, 2020). This gap is not a mere academic curiosity; it also has important practical and scientific consequences. As **AI** systems become increasingly embedded in decision-making processes, from medical diagnosis to other high-impact applications, it becomes essential to quantify the uncertainty associated with their predictions and to establish under which conditions those predictions can be considered reliable. At the same time, a deeper understanding of how neural networks process information is necessary to improve efficiency, guide architectural design, and facilitate the transfer of learned models to related tasks, thereby reducing the need to retrain large networks from scratch. In the broader scientific domain, **AI** is increasingly used not only to analyse data but also to extract regularities, generate hypotheses,

and support the discovery of new laws from complex datasets (Carleo et al., 2019; Mehta et al., 2019; Romera-Paredes et al., 2024; Davies et al., 2021). However, this ambition sits in tension with the limited interpretability of modern learning systems: when the internal logic underlying their conclusions remains opaque, it becomes difficult to assess the nature of the knowledge they capture and the extent to which their outputs contribute to scientific understanding.

These difficulties motivate some of the central open questions in neural network theory (Zdeborová, 2020; Goodfellow et al., 2016; Mehta et al., 2019). Why do highly overparameterised models usually generalise well rather than overfit the training data? Which properties of the data, the network architecture, and the optimisation algorithm determine the effective complexity of the learned solution? Why does gradient-based learning, despite being applied in extremely high-dimensional and non-convex spaces, so often converge to models with strong predictive performance? More broadly, how are internal representations organised across layers, which features of the loss landscape shape the learning dynamics, and how are robustness, uncertainty, and interpretability related? Dealing with these questions is essential for turning the empirical success of modern AI into a more principled understanding of learning systems.

At this point, a natural question arises: why should a statistical physicist care about these problems? The answer is that many of the difficulties posed by modern learning systems are, at their core, problems of collective behaviour in high-dimensional spaces (Mehta et al., 2019; A. Engel et al., 2001). Neural networks are composed of many interacting degrees of freedom; their training can be viewed as a dynamical exploration of a complex (loss) landscape; and their characteristic properties, such as generalisation, robustness, and sensitivity to perturbations, emerge collectively rather than being straightforwardly reducible to the action of individual parameters. In this respect, the challenge is closely related to those traditionally addressed by statistical physics, whose aim is precisely to connect microscopic interactions with macroscopic behaviour in large, heterogeneous systems, whose configurational space is only partially accessible in computation. This connection is not new: early studies of the perceptron and the Hopfield network, together with analyses of large random and disordered neural systems, established an important bridge between neural computation and the statistical mechanics of complex systems (Rosenblatt, 1958; Hopfield, 1982; A. Engel et al., 2001). These works showed that learning, memory, and information propagation could be studied in terms of interacting variables, collective phases, and structured high-dimensional landscapes. Although they were developed for shallow networks, they laid the foundations for a statistical-physics perspective that has recovered central relevance in the era of deep learning (Zdeborová, 2020; Mehta et al., 2019).

From this standpoint, neural networks can be studied not only through their predictive performance, but also through the structure of the space in which learning takes place (Mehta et al., 2019; A. Engel et al., 2001). Training defines a dynamics in

a high-dimensional landscape of parameters, while the resulting behaviour reflects the interplay between optimisation, architecture, and data. Questions about generalisation, robustness, and transferability can therefore be reformulated in terms of the geometry of this landscape, the organisation of its accessible regions, and the effective degrees of freedom that control the system's behaviour. In this framework, notions such as configurational space, entropy, and reduced representation become natural analytical tools. This is the perspective adopted in the present thesis, where neural networks are treated as complex systems and studied through the structure of their configuration space and the possibility of identifying reduced yet informative descriptions of their organisation. This chapter first introduces the formal framework of feed-forward neural networks and supervised learning; it then presents the two statistical physics tools that will be used throughout the remainder of this thesis: the Wang-Landau algorithm, for the exploration of configuration space, and Mapping Entropy, as a measure of the information loss inherent in reduced representations.

6.1 Feed-Forward Neural Networks and the Learning Problem

6.1.1 Feed-Forward Neural Networks

The origins of artificial neural networks can be traced back to the late 1950s, when Frank Rosenblatt introduced the perceptron, the simplest feed-forward neural network, conceived as an abstract model of a neuron with a binary response (McCulloch et al., 1943; Rosenblatt, 1958). Over time, this elementary unit was generalised and combined into increasingly expressive architectures (Goodfellow et al., 2016; Bishop, 2006). The simplest type of deep feed-forward model is the fully connected network, also known as the multi-layer perceptron. In this model, the neurons are arranged in successive layers, with each unit in one layer connected to all units in neighbouring layers. (Goodfellow et al., 2016).

Formally, a feed-forward neural network defines a parametrised mapping $f_{\theta} :$

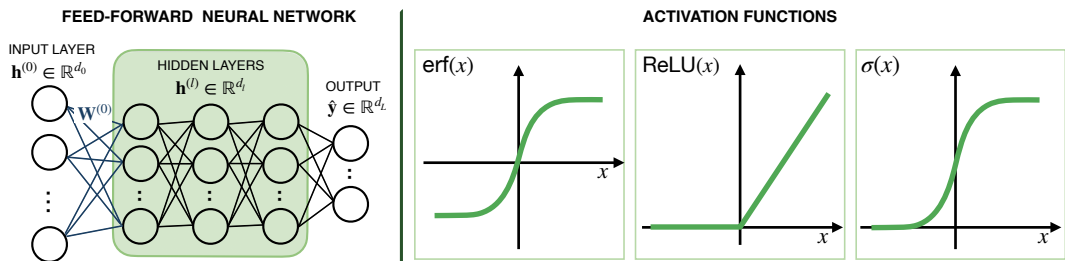


Figure 6.1: Schematic representation of a fully connected feed-forward neural network. The network maps an input layer $\mathbf{h}^{(0)} \in \mathbb{R}^{d_0}$ to an output $\hat{\mathbf{y}} \in \mathbb{R}^{d_L}$ through a sequence of hidden layers $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$ connected by weight matrices $\mathbf{W}^{(l)}$. On the right, examples of commonly used activation functions are shown: the error function $\text{erf}(x)$, the rectified linear unit $\text{ReLU}(x)$, and the sigmoid $\sigma(x)$.

$\mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_L}$ from an input vector $\mathbf{x} \in \mathbb{R}^{d_0}$ to an output $\hat{\mathbf{y}} = f_\theta(\mathbf{x}) \in \mathbb{R}^{d_L}$ through the successive composition of affine transformations and nonlinear activation functions (Goodfellow et al., 2016; Bishop, 2006; Rumelhart et al., 1986). Denoting by $\mathbf{h}^{(\ell)} \in \mathbb{R}^{d_\ell}$ the activation of layer ℓ , with $\mathbf{h}^{(0)} = \mathbf{x}$, the forward propagation through the network is written recursively as

$$\mathbf{h}^{(\ell+1)} = \phi^{(\ell)} \left(W^{(\ell)} \mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)} \right), \quad \ell = 0, \dots, L-1, \quad (6.1)$$

where $W^{(\ell)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ and $\mathbf{b}^{(\ell)} \in \mathbb{R}^{d_{\ell+1}}$ are, respectively, the weight matrix and bias vector of layer ℓ , while $\phi^{(\ell)}$ denotes the corresponding activation function. The set of all trainable parameters,

$$\theta = \left\{ W^{(\ell)}, \mathbf{b}^{(\ell)} \right\}_{\ell=0}^{L-1}, \quad (6.2)$$

fully specifies the network. For a fixed architecture, each choice of θ corresponds to one configuration of the model in parameter space. A schematic representation of a fully connected feed-forward architecture, together with examples of commonly used activation functions, is shown in Figure 6.1.

The presence of nonlinear activation functions is essential. If all $\phi^{(\ell)}$ were linear, the composition of the layers would collapse into a single affine transformation, independently of the network depth (Goodfellow et al., 2016; Bishop, 2006). It is precisely the nonlinear response of the hidden units that allows the model to construct increasingly elaborate internal representations and to approximate highly non-trivial input-output relations. Different choices of activation function, such as sigmoidal functions, hyperbolic tangents, or piecewise-linear activations like the rectified linear unit, lead to distinct geometric and dynamical properties. Still, they all serve a common goal: preventing the network from reducing to a purely linear map (Goodfellow et al., 2016; Bishop, 2006; Nair et al., 2010; Glorot et al., 2011).

Thanks to this non-linearity, feed-forward neural networks are highly expressive; indeed, the universal approximation theorem states that, under suitable assumptions, a network with one hidden layer and sufficiently many units can approximate any continuous function on a compact domain arbitrarily well (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989).

6.1.2 Supervised Learning, Training Error, and Generalisation

In the supervised-learning setting, the goal is to infer a mapping between inputs and outputs from a finite set of examples (Hastie et al., 2009; Shalev-Shwartz et al., 2014; Bishop, 2006). More precisely, given a dataset

$$\mathcal{D} = \{(\mathbf{x}_\mu, \mathbf{y}_\mu)\}_{\mu=1}^M, \quad (6.3)$$

where each input $\mathbf{x}_\mu \in \mathbb{R}^{d_0}$ is associated with a target output $\mathbf{y}_\mu \in \mathbb{R}^{d_L}$, the task is to determine the parameters θ of the network such that the predicted output $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$

approximates the target $\mathbf{y}$ for inputs drawn from the same underlying distribution.

To quantify the quality of the predictions, one introduces a loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$, which measures the discrepancy between the predicted ($\hat{\mathbf{y}}$) and true ($\mathbf{y}$) outputs. The learning problem is then formulated as the minimisation of the empirical risk (Vapnik, 2013; Vapnik, 1999; Hastie et al., 2009),

$$\mathcal{L}(\theta) = \frac{1}{M} \sum_{\mu=1}^M \ell(f_{\theta}(\mathbf{x}_{\mu}), \mathbf{y}_{\mu}), \quad (6.4)$$

which defines an objective function over the parameter space. In this way, training a neural network corresponds to solving a high-dimensional optimisation problem, where each configuration θ is associated with a loss value.

The empirical risk $\mathcal{L}(\theta)$ measures the average loss over the training dataset and is therefore often referred to as the training error (Bishop, 2006; Hastie et al., 2009). However, the ultimate objective of learning is not merely to minimise the error on the observed data, but to achieve good performance on new, unseen samples drawn from the same underlying distribution. To this end, one typically distinguishes between a training set, used to fit the parameters, and a test set, consisting of data not used during training, which provides an estimate of the *generalisation* error (Bishop, 2006; Hastie et al., 2009; Shalev-Shwartz et al., 2014).

Formally, if the data are assumed to be drawn from an unknown distribution $P(\mathbf{x}, \mathbf{y})$, the generalisation error can be defined as the expected loss (Vapnik, 2013; Shalev-Shwartz et al., 2014),

$$\mathcal{L}_{\text{gen}}(\theta) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P} [\ell(f_{\theta}(\mathbf{x}), \mathbf{y})], \quad (6.5)$$

which is approximated in practice by evaluating the model on a test dataset. The difference between training and generalisation error lies at the core of the learning problem: while the former can always be reduced by fitting the training data more closely, the latter reflects the ability of the model to capture the underlying structure of the data rather than dataset-specific, possibly spurious correlations (Vapnik, 2013; Hastie et al., 2009; Shalev-Shwartz et al., 2014).

6.1.3 Gradient-Based Optimisation

In practice, the empirical risk $\mathcal{L}(\theta)$ is minimised through iterative optimisation algorithms that update the parameters in the local direction of decreasing loss (Goodfellow et al., 2016; Bottou, 2012; Bengio, 2012). The most basic scheme is gradient descent, in which the parameters evolve according to

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_t), \quad (6.6)$$

where $\eta > 0$ is the learning rate and $\nabla_{\theta} \mathcal{L}(\theta)$ denotes the gradient of the loss with respect to the parameters.

The efficient evaluation of this gradient is made possible by backpropagation, which exploits the layered structure of the network to compute derivatives recursively via the chain rule (Linnainmaa, 1976; Werbos, 1982; Rumelhart et al., 1986). Since the network output is obtained as a composition of layer-wise transformations, the derivative of the loss with respect to the parameters $\theta^{(\ell)} = \{W^{(\ell)}, \mathbf{b}^{(\ell)}\}$ of layer ℓ can be written schematically as

$$\frac{\partial \mathcal{L}}{\partial \theta^{(\ell)}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{h}^{(L-1)}} \cdots \frac{\partial \mathbf{h}^{(\ell+2)}}{\partial \mathbf{h}^{(\ell+1)}} \frac{\partial \mathbf{h}^{(\ell+1)}}{\partial \theta^{(\ell)}}. \quad (6.7)$$

After the output $\hat{\mathbf{y}} = f_{\theta}(\mathbf{x})$ and the loss $\mathcal{L}$ have been computed by forward propagation, one first evaluates $\partial \mathcal{L} / \partial \hat{\mathbf{y}}$, which measures how a small change in the prediction affects the loss. This quantity is then propagated backwards from one layer to the previous one, so that the sensitivity of the loss to each intermediate activation can be reused to compute the gradients with respect to the corresponding weights and biases. In this way, all derivatives are obtained exactly, but without repeatedly recomputing the same intermediate terms. As a result, the overall computational cost remains the same order of magnitude to that of a single forward propagation, making the training of large networks computationally feasible (Rumelhart et al., 1986; Goodfellow et al., 2016).

In large-scale settings, however, evaluating the full gradient over the entire dataset at every update step is computationally prohibitive and often unnecessary. For this reason, stochastic gradient descent and its variants are typically employed (Robbins et al., 1951; Bottou, 2012; Goodfellow et al., 2016). In this model, the gradient is estimated from a subset of the data, known as a mini-batch. If $\mathcal{B}_t$ denotes the mini-batch used at iteration t , the update becomes

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}_{\mathcal{B}_t}(\theta_t), \quad (6.8)$$

where

$$\mathcal{L}_{\mathcal{B}_t}(\theta) = \frac{1}{|\mathcal{B}_t|} \sum_{\mu \in \mathcal{B}_t} \ell(f_{\theta}(\mathbf{x}_{\mu}), \mathbf{y}_{\mu}). \quad (6.9)$$

When mini-batches are sampled uniformly from the training set, the corresponding gradient provides an unbiased estimator of the full gradient, but with a finite variance (Bottou, 2012; Shalev-Shwartz et al., 2014). The update direction is therefore only an approximation of the deterministic one and introduces an intrinsic source of stochasticity into the learning dynamics. This stochasticity reduces the computational cost of each update and, at the same time, causes the trajectory in parameter space to fluctuate rather than follow a smooth path toward a minimum. As a consequence, stochastic gradient descent does not simply approximate full-batch gradient descent, but defines a different dynamical process, whose prop-

erties can affect both convergence and generalisation (Bottou, 2012; Mehta et al., 2019). In particular, the noise induced by mini-batch sampling may help the dynamics avoid narrow or unstable regions of the loss landscape and is often argued to favour broader, flatter solutions, which are commonly associated with improved robustness and better generalisation on unseen data (Keskar et al., 2017; Hochreiter et al., 1997b). The training process can therefore be viewed as a trajectory in parameter space, shaped jointly by the geometry of the loss landscape, the statistical structure of the data, and the specific properties of the optimisation algorithm (Mehta et al., 2019; Goodfellow et al., 2016).

6.1.4 The Teacher-Student Setting

A particularly convenient framework for studying learning under controlled conditions is the **Teacher-Student (TS)** setting (Watkin et al., 1993; A. Engel et al., 2001). The basic idea is straightforward: rather than considering data generated by an unknown real-world process, one assumes that the labels are produced by a known model, dubbed the *teacher*. A second model, referred to as the *student*, is then trained on a finite set of examples generated by the teacher, and its performance is assessed by comparing its predictions with the teacher's rule on previously unseen inputs.

In the simplest version of this setting, both teacher and student are perceptrons (Seung et al., 1992; Watkin et al., 1993). The teacher is specified at the beginning by a weight vector $\mathbf{w}^*$, typically sampled from a prescribed distribution and then kept fixed. Given an input $\mathbf{x} \in \mathbb{R}^{d_0}$ drawn from a chosen distribution, the teacher assigns the label

$$y = \phi(\mathbf{w}^* \cdot \mathbf{x}), \quad (6.10)$$

where ϕ is a fixed activation function. Repeating this construction for M independent inputs yields the training set

$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^M. \quad (6.11)$$

The student is a second perceptron, with weight vector $\mathbf{w}$, that is adjusted using the examples in $\mathcal{D}_{\text{train}}$.

The same construction can be extended in a natural way to deeper architectures (Saad et al., 1995; Goldt et al., 2019; Goldt et al., 2020). One may consider a teacher network with parameters θ^* , generating outputs according to

$$\mathbf{y} = f_{\theta^*}(\mathbf{x}), \quad (6.12)$$

and a student network, with parameters θ , trained on examples produced by the teacher. The two networks may have the same architecture, but this is not necessary: one may also study mismatched situations, in which student and teacher differ in size, depth, or activation function. In all cases, the central advantage of

the framework remains the same: the target rule is known exactly, and the effect of architectural choices, finite sampling, and optimisation can be analysed without the additional complications of poorly controlled data (Watkin et al., 1993; A. Engel et al., 2001).

This setting has played a central role in the statistical mechanics of learning since the late 1980s. One of the earliest lines of work concerned the storage problem for the perceptron, where analytical methods made it possible to characterise the typical capacity of a network to implement random input-output associations and to identify sharp transitions between regimes in which all training constraints can or cannot be satisfied simultaneously. These results established a prototype for the study of neural networks as disordered systems (Gardner, 1988; Gardner et al., 1988).

A second major achievement of the TS framework was the analytical study of generalisation. In perceptron models, the generalisation error can be expressed in terms of a small number of macroscopic observables, most notably the overlap between teacher and student weights. This reduction allowed explicit predictions for learning curves, asymptotic generalisation error, and the dependence of performance on the size of the training set (Seung et al., 1992; Watkin et al., 1993). Closely related analyses also clarified the role of different training protocols, including online learning rules, for which exact dynamical equations can be derived in the thermodynamic limit (Biehl et al., 1995).

The same methodology was later extended to multilayer architectures, such as committee machines and soft committee machines. In these models, the TS setting revealed phenomena that are absent in the single-layer case, including hidden-unit specialisation, symmetry breaking, and phase transitions separating qualitatively distinct learning regimes (Saad et al., 1995; A. Engel et al., 2001). More recently, related TS constructions have continued to provide a useful theoretical laboratory for the analysis of modern high-dimensional learning problems, including multiclass classification and overparameterised random-feature models (Goldt et al., 2019; Goldt et al., 2020; Cornacchia et al., 2023; Worschech et al., 2024).

For this reason, the TS setting remains a valuable theoretical framework. Its main strength is that, by fixing the data-generating rule, it allows one to isolate and analyse the effects of architecture, sample size, noise, and optimisation in a controlled way. At the same time, this level of control comes at the price of a highly idealised description of the data, making it necessary to develop complementary approaches able to probe analogous questions in settings where the structure of the problem is not imposed *a priori*.

6.1.5 Beyond Feed-Forward Architectures

The fully connected feed-forward network provides the simplest and most direct realisation of a parametrised mapping between inputs and outputs. However, it

represents only one element within a much broader family of neural network architectures that have been developed to address the specific structure of different data modalities and learning tasks (Goodfellow et al., 2016).

A first important generalisation is provided by convolutional neural networks, which are designed to process data with an underlying spatial organisation, such as images. In these models, the connectivity pattern is no longer fully dense, but constrained by locality and weight sharing, so that the same set of parameters is applied across different regions of the input (LeCun et al., 1998). This inductive bias significantly reduces the number of parameters and allows the network to exploit translational invariances present in the data.

Another major class is given by recurrent neural networks, which are tailored to sequential or time-dependent data (Elman, 1990). In this case, the architecture introduces feedback connections that allow the network to maintain an internal state and to process inputs in a temporally ordered fashion. Variants such as long short-term memory networks and gated recurrent units have been developed to address stability issues and to capture long-range dependencies in the input sequence (Hochreiter et al., 1997c; Cho et al., 2014).

More recently, attention-based architectures and, in particular, transformer models have become central in a wide range of applications (Bahdanau et al., 2015). These models replace explicit recurrence or convolution with attention mechanisms that allow each element of the input to interact directly with all others (Vaswani et al., 2017). This leads to a highly flexible representation, capable of capturing complex, long-range correlations in high-dimensional data.

Despite their architectural differences, these models share a common underlying principle: they define parametrised functions obtained by composing elementary transformations, typically involving affine maps and nonlinear activations, and are trained through variants of gradient-based optimisation. From this perspective, the fully connected feed-forward network introduced above can be regarded as the simplest reference model, in which the role of architecture is reduced to its minimal form.

For this reason, the analysis developed in the following will focus primarily on feed-forward architectures, where the interplay between data, optimisation, and representation can be studied in a controlled setting. At the same time, many of the concepts that will be introduced, such as the structure of the loss landscape or the identification of effective degrees of freedom, are expected to extend beyond this specific class of models and to remain relevant for more complex architectures.

6.2 Wang-Landau Sampling

Wang-Landau (WL) sampling is a generalised-ensemble Monte Carlo method for estimating the density of states of a system. Its main motivation lies in the limitations of conventional importance-sampling schemes: at fixed temperature, canon-

ical simulations efficiently explore only the energy region that gives the dominant contribution to the Boltzmann distribution, and therefore are known to become inefficient in presence of metastability, free-energy barriers, or phase coexistence (Berg et al., 1992; F. Wang et al., 2001b; Mitsutake et al., 2001; D. P. Landau et al., 2004). By targeting the density of states rather than a temperature-dependent equilibrium distribution, WL sampling provides a route to reconstructing thermodynamic properties over a broad temperature range within a single computational framework.

In the following, the method is introduced first through the density of states and its thermodynamic role, then through the flat-histogram idea and the original WL algorithm, and finally through the main practical and theoretical issues associated with its implementation and convergence.

6.2.1 Density of States and Thermodynamic Reconstruction

Building on the statistical mechanics framework introduced in Chapter 2, it is useful to rewrite the partition function in terms of the density of states $g(E)$, namely the number of microscopic configurations associated with a given energy E . Restricting ourselves, for simplicity, to the configurational contribution and to systems with discrete energy levels, one has

$$Z(\beta) = \sum_{\mathbf{x}} e^{-\beta E(\mathbf{x})} = \sum_E g(E) e^{-\beta E}, \quad (6.13)$$

where $\beta = 1/(k_B T)$, T is the temperature, and the second sum runs over the accessible energy levels. This representation separates two contributions to the canonical weight at fixed energy: the Boltzmann factor $e^{-\beta E}$, which suppresses high-energy states, and the density of states $g(E)$, which counts how many configurations realise that energy. The function $g(E)$ therefore encodes the temperature-independent statistical structure of the system, while the temperature dependence enters only through the Boltzmann weight.

The central role of $g(E)$ follows immediately from Equation 6.13. Once the density of states is known, equilibrium thermodynamic quantities can be reconstructed over the whole temperature range without performing independent simulations at each temperature. In particular, the canonical probability of observing energy E is

$$P_\beta(E) = \frac{g(E) e^{-\beta E}}{Z(\beta)}, \quad (6.14)$$

so that thermal averages can be obtained by reweighting over energy. For observables depending only on E , one has

$$\langle A \rangle_\beta = \frac{\sum_E A(E) g(E) e^{-\beta E}}{\sum_E g(E) e^{-\beta E}}. \quad (6.15)$$

Free energy, internal energy, entropy, and response functions then follow from

the standard thermodynamic relations once the partition function has been reconstructed.

The practical difficulty is that $g(E)$ is rarely accessible analytically. Determining it requires a detailed characterisation of how the microscopic configurations are distributed over the energy spectrum, which is feasible only in a limited number of idealised cases and rapidly becomes intractable as the size and complexity of the system increase. For this reason, one typically seeks a statistical estimate of $g(E)$ based on stochastic exploration of configuration space. This is the context in which generalised-ensemble methods become relevant: by modifying the sampling weights, they enable the exploration of broad energy regions and thereby the indirect estimation of the density of states (Berg et al., 1992; Mitsutake et al., 2001; D. P. Landau et al., 2004).

6.2.2 Generalised-Ensemble Motivation and the Flat-Histogram Idea

The difficulty of estimating $g(E)$ is closely connected to a basic limitation of canonical importance sampling. In a standard Monte Carlo simulation at fixed temperature, configurations are generated according to the Boltzmann weight, so that the induced energy distribution is

$$P_\beta(E) \propto g(E)e^{-\beta E}. \quad (6.16)$$

As a consequence, the simulation predominantly explores the energy range that is relevant at the chosen temperature, while energetically or entropically suppressed regions are visited only rarely. This becomes especially problematic in systems with rough free-energy landscapes, metastable states, or phase coexistence, where barriers may strongly hinder transitions between relevant regions of the configuration space.

Generalised-ensemble methods were introduced precisely to mitigate this difficulty. Their common idea is to replace the canonical Boltzmann weight with a modified sampling weight that enhances exploration of states that would otherwise be poorly sampled. Rather than directly generating a physical ensemble at fixed temperature, these methods define an auxiliary stochastic dynamics whose output can later be reweighted to recover equilibrium information (Berg et al., 1992; D. P. Landau et al., 2004).

A particularly important strategy within this class is to aim for a flat exploration in energy space. If configurations could be sampled with a weight

$$w(E) \propto \frac{1}{g(E)}, \quad (6.17)$$

then the resulting energy distribution would satisfy

$$P(E) \propto g(E)w(E) \approx \text{const.}, \quad (6.18)$$

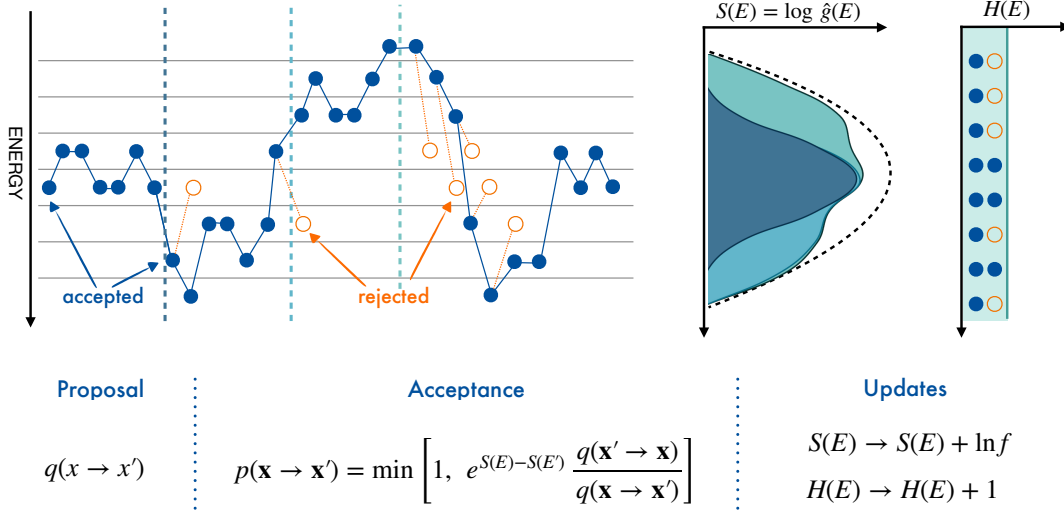


Figure 6.2: Schematic representation of the original *Wang-Landau* algorithm. Starting from a configuration x , a trial move $x \rightarrow x'$ is proposed according to the kernel $q(x \rightarrow x')$ and accepted with probability $p(x \rightarrow x') = \min\left[1, e^{S(E)-S(E')} \frac{q(x' \rightarrow x)}{q(x \rightarrow x')}\right]$, where $S(E) = \log \hat{g}(E)$ is the current estimator of the logarithmic density of states. After each Monte Carlo step, both $S(E)$ and the histogram $H(E)$ are updated at the visited energy according to $S(E) \rightarrow S(E) + \ln f$ and $H(E) \rightarrow H(E) + 1$.

so that all accessible energies would be visited with comparable frequency. In that case, the resulting random walk in energy would no longer remain confined to the narrow interval favoured by a canonical ensemble, but would instead probe the energy range more uniformly. The flat-histogram idea is therefore useful not because a uniform energy distribution is of physical interest in itself, but because it provides an efficient route to estimating the density of states (F. Wang et al., 2001b; D. P. Landau et al., 2004).

The main difficulty is that the ideal weight $w(E) \propto 1/g(E)$ requires prior knowledge of the same quantity one wishes to determine. WL sampling resolves this difficulty by estimating the density of states iteratively during the simulation and using the current estimate to update the sampling bias. It can therefore be viewed as an adaptive realisation of the flat-histogram principle (F. Wang et al., 2001b; F. Wang et al., 2001a; D. P. Landau et al., 2004; Zhou et al., 2005).

6.2.3 The Original Wang-Landau Algorithm

The problem is therefore no longer to identify the target distribution, but to construct a Monte Carlo scheme that approaches it without knowing $g(E)$ in advance. The original WL algorithm addresses this point by introducing a running estimate $\hat{g}(E)$, which is used directly in the acceptance probabilities and progressively refined during the simulation.

More precisely, one defines a random walk in configuration space such that, at any stage of the procedure, the current estimator $\hat{g}(E)$ plays the role that the exact density of states would have in the ideal flat-histogram ensemble. For symmetric

proposal probabilities, a trial move from a configuration of energy E to one of energy E' is therefore accepted with probability

$$p(E \rightarrow E') = \min \left[1, \frac{\hat{g}(E)}{\hat{g}(E')} \right], \quad (6.19)$$

which is the natural Metropolis prescription for an auxiliary ensemble in which the statistical weight of a configuration is proportional to $1/\hat{g}(E)$. If the estimator were held fixed, the corresponding stationary probability of observing energy E would then be

$$P_{\hat{g}}(E) \propto \frac{g(E)}{\hat{g}(E)}. \quad (6.20)$$

Hence, whenever $\hat{g}(E)$ coincides with the true density of states up to a multiplicative constant, the energy distribution becomes flat. The role of the algorithm is precisely to drive the estimator toward this self-consistent solution (F. Wang et al., 2001b; F. Wang et al., 2001a; D. P. Landau et al., 2004; Zhou et al., 2005).

Since no prior information on $g(E)$ is assumed, the procedure can be started from the simplest possible choice, namely a uniform guess $\hat{g}(E) = 1$ over the accessible energy range. In addition, one also keeps track of a histogram $H(E)$, which counts how many times each energy level has been visited during the sampling.

Starting from this initial condition, the simulation proceeds by alternating between the stochastic evolution of the system and the update of the current estimate $\hat{g}(E)$ together with the histogram $H(E)$. After each Monte Carlo step, both quantities are modified at the energy actually visited by the walk. In the original scheme, if that energy is E , one performs the updates

$$\hat{g}(E) \rightarrow \hat{g}(E) f, \quad H(E) \rightarrow H(E) + 1, \quad (6.21)$$

where $f > 1$ is the modification factor. The update is therefore applied to the energy effectively occupied after the accept-reject step, so that both the current estimate of the density of states and the visitation histogram evolve together with the random walk (F. Wang et al., 2001b; F. Wang et al., 2001a; D. P. Landau et al., 2004).

The rationale behind this prescription is to introduce a negative feedback on the exploration of energy space. Energies that are visited frequently are repeatedly assigned a larger estimated density of states and, through Equation 6.19, become progressively less likely to be revisited. Conversely, energies that have been sampled only rarely are penalised less strongly and therefore remain comparatively more accessible. In this way, the coupled evolution of $\hat{g}(E)$ and $H(E)$ progressively compensates sampling imbalances and drives the walk toward the flat-histogram regime (F. Wang et al., 2001b; Zhou et al., 2005).

The parameter f controls the amplitude of this correction. At the beginning of the simulation, when the estimate $\hat{g}(E)$ is still far from the true density of states, f is chosen sufficiently large to allow rapid reshaping of the profile of $\hat{g}(E)$. As the sim-

ulation proceeds and the histogram becomes progressively flatter, finer corrections are required. For this reason, the original WL algorithm does not keep f fixed, but reduces it through a sequence of refinement stages. At each stage, the histogram $H(E)$ is accumulated until it satisfies a prescribed flatness criterion, after which the histogram is reset and the modification factor is reduced.

The refinement is performed according to the recursive rule

$$f_{n+1} = \sqrt{f_n}, \quad (6.22)$$

starting from an initial value typically chosen as $f_0 = e$. The simulation then proceeds through successive iterations, each characterised by a smaller modification factor and therefore by a less impactful update of the density-of-states estimate. The logic is that the early stages should rapidly establish the global shape of $\hat{g}(E)$, whereas the later ones should progressively improve its accuracy.

The condition used to decide when a given refinement stage is complete is usually formulated in terms of the histogram $H(E)$. In the original work, the histogram was considered sufficiently flat when every visited energy satisfied

$$H(E) \geq c \langle H \rangle, \quad (6.23)$$

where $\langle H \rangle$ denotes the average histogram count over the sampled energy range and c is a chosen threshold, often taken between 0.8 and 0.95. Once this condition is met, the current estimate $\hat{g}(E)$ is retained, the histogram is reset, and a new refinement stage begins with a smaller value of f .

The original WL scheme therefore combines three essential ingredients: a transition rule that would generate the desired flat-histogram ensemble if the exact density of states were known, a multiplicative update of the running estimate $\hat{g}(E)$, and a staged reduction of the modification factor f . Together, these elements define a self-consistent procedure in which the density of states is learned during the sampling itself rather than provided in advance (F. Wang et al., 2001b; F. Wang et al., 2001a; D. P. Landau et al., 2004).

6.2.4 Practical Aspects of the Wang–Landau Method

Beyond its formal definition, the practical implementation of WL sampling requires a number of additional choices that influence both the accuracy of the density of states estimate and the efficiency of the exploration. In particular, one must address the numerical representation of the estimator, the treatment of the energy axis – including its discretisation and possible restriction to a finite interval – and the definition of the proposal probabilities entering the Monte Carlo sampling.

A first practical issue concerns the numerical representation of the density of states estimate. In principle, the algorithm can be formulated directly in terms of $\hat{g}(E)$, as in Equation 6.19 and Equation 6.21. In practice, however, this is rarely

convenient, because the density of states typically varies over many orders of magnitude and may increase extremely rapidly with system size. Storing and updating $\hat{g}(E)$ directly can therefore lead to numerical overflows, underflows, or substantial loss of precision.

For this reason, WL implementations are usually formulated in terms of the logarithm of the estimate,

$$S(E) = \ln \hat{g}(E), \quad (6.24)$$

which plays the role of a microcanonical entropy up to an irrelevant additive constant. In this representation, the acceptance probability for symmetric proposal moves becomes

$$p(E \rightarrow E') = \min [1, \exp (S(E) - S(E'))], \quad (6.25)$$

while the multiplicative update of Equation 6.21 is replaced by the additive rule

$$S(E) \rightarrow S(E) + \ln f. \quad (6.26)$$

The modification factor is therefore naturally handled through its logarithm, and the iterative refinement of the density of states is recast as a sequence of additive corrections to $S(E)$ (D. P. Landau et al., 2004; Shell et al., 2002).

This reformulation leaves the logic of the algorithm unchanged, but improves its numerical stability and interpretability. In particular, the indeterminacy of the overall normalisation of $\hat{g}(E)$ corresponds simply to an additive constant in $S(E)$, which does not affect either the sampling dynamics or the reconstruction of thermodynamic observables based on derivatives or relative entropy differences.

A second important practical aspect concerns the representation of the energy axis on which the quantities $\hat{g}(E)$, $S(E)$, and $H(E)$ are defined. For systems with a discrete energy spectrum, these quantities can in principle be associated directly with the allowed energy values. For systems with continuous energies, instead, the algorithm must be implemented on a discretised energy axis, obtained by partitioning the relevant energy range into bins of finite width ΔE . In this case, the method no longer estimates the density of states at exact energy values, but rather a CG approximation associated with the chosen discretisation (Shell et al., 2002; Zhou et al., 2006; Poulain et al., 2006).

The choice of the bin width involves a nontrivial compromise. If ΔE is too large, the resulting estimate loses resolution and may smoothen out relevant features of the density of states. If it is too small, the number of sampled bins increases, the statistics collected in each bin decrease, and the convergence of the algorithm may become significantly slower. The discretisation of the energy axis is therefore not a purely technical detail, but part of the effective definition of the estimated density of states itself.

A related issue concerns the range of energies over which the random walk is performed. In some systems the full energy spectrum is naturally bounded and

known in advance, but in many cases it is either not known exactly, effectively unbounded, or so large that sampling it entirely would be computationally impractical. Moreover, thermodynamic observables are often controlled only by a restricted portion of the spectrum, so that exploring the entire energy axis would be unnecessary. For these reasons, WL simulations are frequently carried out over a finite energy interval chosen to include the region of physical interest.

The selection of this interval requires particular care. If the chosen window is too narrow, energetically relevant configurations may be excluded and the reconstructed thermodynamics may become biased or incomplete. If it is excessively broad, the exploration may become inefficient and the statistical effort may be spread over regions that contribute negligibly to the observables of interest. In practical applications, the sampled interval is therefore typically determined on the basis of prior physical knowledge, preliminary simulations, or iterative refinement.

Once a finite energy window has been selected, the treatment of moves that would bring the system outside of the prescribed interval must also be specified consistently. The most common choice is to reject such moves and to treat the current energy as visited once again, so that both the histogram and the density-of-states estimate are updated at the previously occupied energy. In this way, the random walk remains confined to the selected support while preserving the internal consistency of the update scheme. More elaborate strategies, based for example on overlapping windows, may also be employed when the relevant energy range is too wide to be sampled efficiently in a single simulation (F. Wang et al., 2001b; Belardinelli et al., 2007a).

A final important element of the practical implementation is the choice of the proposal probability governing the attempted moves in configuration space. Denoting by $q(\mathbf{x} \rightarrow \mathbf{x}')$ the probability of proposing a transition from configuration $\mathbf{x}$ to configuration $\mathbf{x}'$, the acceptance rule of Equation 6.19 applies directly only when the proposal mechanism is symmetric, namely when

$$q(\mathbf{x} \rightarrow \mathbf{x}') = q(\mathbf{x}' \rightarrow \mathbf{x}). \quad (6.27)$$

In many practical situations, however, restricting the dynamics to symmetric proposals would be unnecessarily limiting. The proposal mechanism may need to be adapted to the structure of the configuration space, to the presence of geometric or dynamical constraints, or more simply to improve the efficiency of the exploration by favouring moves of a given type or scale. In such cases, the proposal probabilities are generally asymmetric, and this asymmetry must be taken into account explicitly in order to preserve the intended sampled distribution (Hastings, 1970; Shell et al., 2002).

More specifically, if a move from $\mathbf{x}$ to $\mathbf{x}'$ changes the energy from E to E' , the

correct WL acceptance probability is given by the Metropolis-Hastings form

$$p(\mathbf{x} \rightarrow \mathbf{x}') = \min \left[1, \frac{\hat{g}(E) q(\mathbf{x}' \rightarrow \mathbf{x})}{\hat{g}(E') q(\mathbf{x} \rightarrow \mathbf{x}')} \right]. \quad (6.28)$$

Equivalently, in terms of the logarithmic estimator introduced above, this can be written as

$$p(\mathbf{x} \rightarrow \mathbf{x}') = \min \left[1, \exp \left(S(E) - S(E') + \ln \frac{q(\mathbf{x}' \rightarrow \mathbf{x})}{q(\mathbf{x} \rightarrow \mathbf{x}')} \right) \right]. \quad (6.29)$$

The symmetric rule of Equation 6.25 is therefore recovered as the special case in which the proposal ratio is equal to unity.

The choice of the proposal dynamics does not alter the target density of states, but it can strongly affect the efficiency with which the energy landscape is explored. Different proposal mechanisms may lead to substantially different tunnelling rates across entropic barriers, different round-trip times between low- and high-energy regions, and therefore different practical convergence properties. For this reason, the definition of the proposal probability is not merely an auxiliary technical detail, but an important component of the overall effectiveness of a WL simulation.

In the following, the expression “density of states estimate” will be used generically to refer either to $\hat{g}(E)$ itself or, when the logarithmic representation is adopted, to the corresponding quantity $S(E) = \ln \hat{g}(E)$.

6.2.5 Error Saturation in the Original Wang–Landau Scheme

The histogram-flatness criterion of the original WL procedure is an operational rule for deciding when a refinement stage should be terminated and the modification factor reduced. However, it does not directly control the actual error of the estimated density of states (Zhou et al., 2005; Belardinelli et al., 2007b; Zhou et al., 2008). To make this distinction explicit, it is useful to introduce few diagnostic quantities. Denoting by $g(E, t)$ the estimated density of states at Monte Carlo time t , and by $g_{\text{ex}}(E)$ the exact one, the local errors are defined as

$$\epsilon(E, t) = \left| 1 - \frac{\ln g(E, t)}{\ln g_{\text{ex}}(E)} \right|, \quad \eta(E, t) = \left| 1 - \frac{g(E, t)}{g_{\text{ex}}(E)} \right|. \quad (6.30)$$

Their averages over the sampled energy levels provide two global measures of the quality of the estimate,

$$\langle \epsilon(t) \rangle = \frac{1}{N-1} \sum_E \epsilon(E, t), \quad \langle \eta(t) \rangle = \frac{1}{N-1} \sum_E \eta(E, t), \quad (6.31)$$

where N is the number of sampled energy values or bins (Belardinelli et al., 2007a). In parallel, the flatness of the visitation histogram can be monitored through global quantities constructed from $H(E, t)$, which denotes the number of visits accumu-

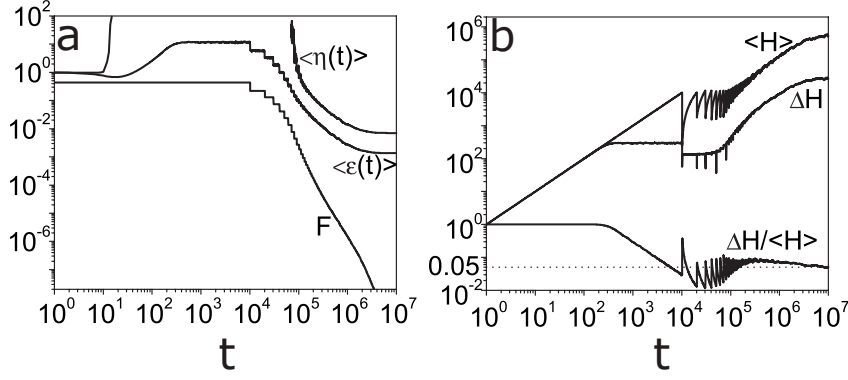


Figure 6.3: Dynamical behaviour of the original WL algorithm. Panel (a) reports the refinement parameter $F = \ln f$ together with the mean error measures $\langle \epsilon(t) \rangle$ and $\langle \eta(t) \rangle$. Panel (b) shows the average histogram height $\langle H(t) \rangle$, its width $\Delta H(t)$, and the relative fluctuation $\Delta H(t)/\langle H(t) \rangle$. Figure reproduced under license from (Belardinelli et al., 2007a)

lated up to time t in the energy level or bin E . In particular, one may consider its average height

$$\langle H(t) \rangle = \frac{1}{N} \sum_E H(E, t), \quad (6.32)$$

its width

$$\Delta H(t) = H_{\max}(t) - H_{\min}(t), \quad (6.33)$$

and the relative fluctuation $\Delta H(t)/\langle H(t) \rangle$.

A first limitation of the original scheme is that its late-stage refinement becomes progressively inefficient. Writing $F_k = \ln f_k$, the original updating rule $f_{k+1} = \sqrt{f_k}$ becomes

$$F_{k+1} = \frac{F_k}{2}. \quad (6.34)$$

Since each visit changes the logarithmic estimate $S(E)$ by an amount F_k (Equation 6.26), the individual corrections become exponentially smaller from one refinement stage to the next. Consequently, obtaining an overall correction of comparable size requires an exponentially larger number of visits. This behaviour is already visible in Figure 6.3, where the successive reductions of F are accompanied by progressively smaller decreases in $\langle \epsilon(t) \rangle$ and $\langle \eta(t) \rangle$.

This point can be made more quantitative. For fixed refinement parameter $F = \ln f$, the residual statistical error scales as

$$\delta S \sim \sqrt{F}, \quad (6.35)$$

so reducing F does improve the estimate, but only slowly: each successive refinement stage yields a smaller gain than the previous one. In this sense, the original schedule suffers from diminishing efficiency even before one addresses its true asymptotic behaviour (Zhou et al., 2005; Belardinelli et al., 2007b).

The more serious problem is that, under the original refinement schedule, the

cumulative strength of all future updates remains finite. Indeed, during the k -th refinement stage, each update changes $S(E)$ by an amount of order F_k . Because the original scheme imposes the geometric decay $F_{k+1} = F_k/2$, the cumulative amplitude of all corrections that remain available after a given stage k_0 is bounded by

$$\sum_{k=k_0}^{\infty} F_k = F_{k_0} \sum_{n=0}^{\infty} 2^{-n} = 2F_{k_0}. \quad (6.36)$$

This quantity is finite for any k_0 . Consequently, once the residual stochastic fluctuations of the estimate become larger than the total corrective power still available at later stages, the algorithm can no longer remove them. The estimate then ceases to improve asymptotically and the error approaches a non-zero plateau (see [Figure 6.3](#)). This is the origin of the saturation effect in the original WL scheme ([Belardinelli et al., 2007b](#)).

The same conclusion is reflected in the histogram itself. As shown in [Figure 6.3](#), the relative fluctuation $\Delta H/\langle H \rangle$ decreases at first and satisfies the chosen flatness threshold, but both $\langle H \rangle$ and ΔH eventually approach regimes in which they only fluctuate around nearly constant mean values, while the error measures $\langle \epsilon(t) \rangle$ and $\langle \eta(t) \rangle$ no longer decay. Histogram flatness is therefore operationally useful, but it is not a sufficient condition for asymptotic convergence of the density-of-states estimate ([Zhou et al., 2005](#); [Belardinelli et al., 2007b](#); [Zhou et al., 2008](#)).

The limitation of the original WL procedure is thus not merely that convergence becomes slow at late times, but that the refinement schedule itself prevents the error from vanishing asymptotically. This is the central reason for replacing the geometric reduction of F by a slower time-dependent law, leading to the $1/t$ algorithm discussed in the next section.

6.2.6 The $1/t$ Algorithm

Since the saturation of the original WL scheme originates from the fact that the cumulative strength of all future updates remains finite, the natural remedy is to adopt a refinement schedule that decays more slowly in time. More generally, one may consider a time-dependent refinement parameter of the form

$$F(t) \propto t^{-p}, \quad (6.37)$$

with t the Monte Carlo time and $p > 0$. The key requirement is that the cumulative correction

$$\sum_t F(t), \quad (6.38)$$

must remain divergent, so that the algorithm retains an unbounded capacity to correct the estimate at late times. This excludes schedules that decay too rapidly, such as the geometric refinement of the original WL procedure. Within the family de-

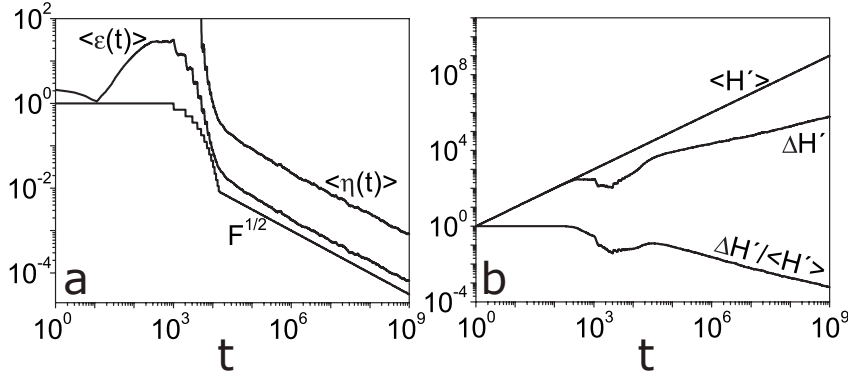


Figure 6.4: Dynamical behaviour of the $1/t$ refinement scheme. Panel (a) shows the decay of the error measure $\langle \epsilon(t) \rangle$, together with the fluctuation measure $\langle \eta(t) \rangle$; the reference slope $t^{-1/2}$ is also reported. Panel (b) shows the average histogram height $\langle H' \rangle$, its width $\Delta H'$, and the relative fluctuation $\Delta H' / \langle H' \rangle$. Figure reproduced under license from (Belardinelli et al., 2007a)

finied in Equation 6.37, the optimal choice is the marginal case

$$F(t) = \frac{1}{t}, \quad (6.39)$$

which is the slowest decay still ensuring that the update amplitude vanishes asymptotically. Faster decays with $p > 1$ again lead to a finite total correction and therefore to non-convergence, whereas slower decays with $p < 1$ preserve convergence but reduce the error more slowly. The case $p = 1$ therefore provides the best compromise between continued correction and asymptotic refinement (Belardinelli et al., 2007a; Belardinelli et al., 2007b; Belardinelli et al., 2014).

The practical implementation of the $1/t$ algorithm retains the initial stage of the original WL scheme, during which the estimate is rapidly shaped through successive reductions of the modification factor, following the standard exponential reduction schedule (Equation 6.34). The crucial change concerns the late-time regime: once the refinement parameter becomes of the order of the inverse Monte Carlo time,

$$F(t) \lesssim \frac{1}{t}, \quad (6.40)$$

the flatness-controlled iterations are abandoned and the simulation enters a continuous refinement stage in which

$$F(t) = \frac{1}{t} \quad \text{for } t \geq t_c, \quad (6.41)$$

with t_c the crossover time. From that point onward, the logarithmic density-of-states estimate is updated as

$$S(E) \rightarrow S(E) + \frac{1}{t}, \quad (6.42)$$

whenever the random walk visits the energy level or bin E . In this regime, the ex-

plicit histogram-flatness test is no longer needed to determine the refinement schedule, since the update amplitude is controlled directly by time (Belardinelli et al., 2007a; Belardinelli et al., 2007b).

The advantage of this choice is twofold. First, because

$$\sum_{t=t_c}^{\infty} \frac{1}{t}, \quad (6.43)$$

diverges, the algorithm never exhausts its capacity to correct the estimate, and the saturation mechanism of the original scheme is avoided. Second, the update amplitude still tends to zero, so the stochastic fluctuations of the estimate are progressively reduced. As a result, the error no longer approaches a non-zero plateau, but decays asymptotically as

$$\delta S(t) \sim t^{-1/2}, \quad (6.44)$$

as illustrated in Figure 6.4, where the decay of $\langle \epsilon(t) \rangle$ follows the $t^{-1/2}$ behaviour once the $1/t$ regime is established (Belardinelli et al., 2007a; Belardinelli et al., 2014).

The same figure also shows that, in contrast with the original refinement scheme, the histogram diagnostics continue to evolve consistently with ongoing refinement: the fluctuation measure $\langle \eta(t) \rangle$ keeps decreasing, while the relative histogram width $\Delta H' / \langle H' \rangle$ does not approach the stationary plateau characteristic of error saturation. The $1/t$ prescription therefore restores asymptotic convergence not only formally, through the divergence of Equation 6.43, but also at the level of the observed long-time behaviour of both the error and the histogram fluctuations (Belardinelli et al., 2007a).

The significance of the $1/t$ prescription is that it identifies, and leverages, the threshold between two qualitatively different behaviours: if the refinement parameter decreases faster than $1/t$, the remaining total correction is finite and the error cannot vanish asymptotically; if it decreases more slowly, convergence is preserved but the reduction of the fluctuations is suboptimal. The $1/t$ law is thus the marginal choice that restores asymptotic convergence while maintaining the fastest decay of the error within this class of time-dependent schedules (Belardinelli et al., 2007a; Belardinelli et al., 2014).

6.3 Applications and Broader Scope of Wang-Landau Sampling

Over the last two decades, WL sampling and its variants have been used to address a broad range of statistical mechanics problems in which conventional canonical sampling becomes inefficient because relevant regions of configuration space are separated by strong entropic or free energy barriers. A particularly important class of applications concerns first-order transitions and phase coexistence. In these cases,

the canonical probability distribution is strongly suppressed in the intermediate region between competing phases, so that direct simulations may remain trapped for long times in one metastable basin. By compensating this suppression through an estimate of the density of states, WL sampling enables repeated crossings of the coexistence region and allows one to reconstruct the associated free energy barriers and finite-size thermodynamics with high accuracy (F. Wang et al., 2001b; D. P. Landau et al., 2004; Berg et al., 1992; Belardinelli et al., 2007b).

The usefulness of the method, however, is not restricted to the reconstruction of a one-dimensional density of states $g(E)$. In many systems, energy alone is not sufficient to distinguish between physically relevant configurations or transition pathways. In these cases, the method can be extended to estimate a joint density of states $g(E, q)$, where q denotes an additional collective variable or order parameter introduced to resolve structural, magnetic, or geometrical features that are not uniquely identified by the energy alone. This extension has proved particularly valuable in the study of phase transitions and in systems such as polymers and biomolecules, where energy-degenerate configurations may nevertheless belong to distinct macroscopic states (Zhou et al., 2006; Poulain et al., 2006; Menichetti et al., 2021).

Although lattice and other discrete-state systems remain the most natural setting for the method, WL sampling is not restricted to them. The approach has also been extended to off-lattice and continuous systems, provided that the relevant variables are discretised appropriately. In that case, the estimated density of states must be understood in a CG sense, associated with finite bins in energy or in a multidimensional space of collective variables. Moreover, in continuum models, one must define the target quantity carefully, so as to separate configurational and kinetic contributions consistently and to derive the correct acceptance probabilities for the biased random walk (Shell et al., 2002; Zhou et al., 2006; Poulain et al., 2006). These additional ingredients do not alter the logic of the method, but they make its practical implementation more delicate than in discrete models.

The continuum case also makes clear a more general point: the success of WL sampling depends not only on the refinement schedule, but also on the adequacy of the chosen reaction coordinates. When the slow degrees of freedom of the system are not well resolved by energy alone, a one-dimensional density of states $g(E)$ may be insufficient to capture the relevant barriers. In such cases, extending the method to a multidimensional density of states is not merely a refinement, but often a necessity for obtaining an accurate description of the explored configuration space (Poulain et al., 2006; Zhou et al., 2006). The flexibility of WL sampling lies precisely in this possibility: the method does not require a fixed thermodynamic ensemble, but can be reformulated in terms of generalised densities of states adapted to the structure of the problem under study.

Taken together, these developments show that WL sampling is a remarkably powerful framework for the exploration of complex energy landscapes. Its strength lies in the possibility of reconstructing thermodynamic and structural information

over broad regions of configuration space within a single simulation strategy, while overcoming the trapping problems that affect conventional importance-sampling methods. At the same time, the discussion of the previous sections has made clear that this power comes with important practical and theoretical constraints: the numerical representation of the density of states, the choice of the sampled range and proposal dynamics, the treatment of continuous variables, and the convergence properties of the refinement schedule all play a decisive role in the reliability of the final estimate.

In this perspective, the methodology developed in this thesis naturally fits within the broader scope of WL sampling while also extending it into a direction that, to the best of our knowledge, has not previously been explored in a systematic way, namely the study of neural network configuration spaces. As discussed in the previous section, these systems are characterised by high-dimensional parameter spaces, a strong combinatorial proliferation of admissible configurations, and the presence of broad regions of near-degenerate solutions separated by nontrivial barriers. These are precisely the conditions under which a flat-histogram strategy becomes especially valuable. Rather than focusing only on a single optimised configuration, WL sampling provides a framework to probe the global organisation of the landscape, quantify the abundance of solutions at different effective energies, and access regions that would remain strongly suppressed under conventional training dynamics. For this reason, it provides a particularly well-suited conceptual and computational basis for the analyses presented in the following chapters.

6.4 Mapping Entropy

The problem of reduced representations appears in different forms throughout this thesis. In molecular systems, it arises in the construction of CG descriptions of high-dimensional configurational ensembles; in neural networks, it emerges when one aims to eliminate subsets of neurons, in particular hidden units, so as to decrease model complexity without compromising performance on the task. In both cases, the central issue is the same: a reduced description is obtained through a mapping that lowers the effective dimensionality of the system, and this operation inevitably induces a loss of information.

A quantitative treatment of this loss is provided by the Mapping Entropy (ME), an information-theoretic quantity originally introduced in the context of bottom-up CG of biomolecules (Shell, 2008; Chaimovich et al., 2011; Giulini et al., 2020). Although its historical derivation is rooted in statistical mechanics and in the relative entropy formulation of soft matter CG modelling, the resulting concept is more general: it provides a measure of the information that is irreversibly discarded when a detailed description is replaced by a reduced one. For this reason, it is particularly well suited for the analyses developed in the following chapters, where the same logic will be transferred from molecular mappings to the selection of hidden

neurons in neural networks.

In the following, the **ME** is introduced by first recalling the relative-entropy framework in which it was originally derived, and then isolating the contribution that depends on the mapping alone. This route is useful not only for historical reasons but also because it places the construction on a precise statistical-mechanical basis and clarifies the relation between **CG** representations, reconstructed distributions, and information loss (Giulini et al., 2020; Giulini, 2022).

6.4.1 Relative Entropy Framework

As discussed more in detail in **Chapter 2**, the exact **CG** description of a system is formally encoded in the **Many-body Potential of Mean Force (MB-PMF)**. As already noted, however, this quantity is, in general, not amenable to direct use in simulations, due to its high-dimensional and strongly nontrivial structure. This motivates the introduction of approximate reduced models. Within the bottom-up perspective recalled in **Chapter 2**, one of the most systematic approaches is the *relative entropy* method, which formulates **CG** as a variational problem and introduces a functional to quantify the discrepancy between a detailed reference model and an approximate reduced one.

In its original formulation (Shell, 2008), the relative-entropy approach considers two probability distributions: a target distribution, denoted by p_T , which represents the detailed reference system, and a model distribution, denoted by p_M , which represents an approximate description of the same system. Both distributions are defined over the same configurational space Ω , whose generic element is denoted here by $\mathbf{r}$. Their discrepancy can therefore be directly compared and quantified in a statistically well-defined manner. In this setting, the natural functional measuring the discrepancy between the two descriptions is the **KL divergence** (Kullback et al., 1951),

$$S_{\text{rel}} = \int_{\Omega} d\mathbf{r} p_T(\mathbf{r}) \ln \frac{p_T(\mathbf{r})}{p_M(\mathbf{r})}. \quad (6.45)$$

This quantity is non-negative and vanishes only when the two distributions coincide almost everywhere; it therefore provides a natural criterion to optimise the approximate model (Kullback et al., 1951) by means of variational approaches.

This definition applies when the target and model distributions are defined on the same configurational space. In the **CG** setting, however, this condition is no longer satisfied. One wishes, in fact, to compare the microscopic reference distribution $p_r(\mathbf{r}) \propto e^{-\beta u(\mathbf{r})}$, defined on the full space Ω , with a **CG** model distribution $P_R(\mathbf{R} \mid U) \propto e^{-\beta U(\mathbf{R})}$, defined on the lower-dimensional space Ω_R , where $U(\mathbf{R})$ denotes a generic effective **CG** model and $P_R(\mathbf{R} \mid U)$ the corresponding approximate reduced distribution, to be distinguished from the exact mapped distribution $P_R(\mathbf{R})$ introduced in **Chapter 2**, built on the **MB-PMF**. Since the two distributions live on spaces of different dimensionality, their comparison requires an additional

construction. Specifically, the CG distribution must be transported back to the microscopic configurational space, so as to assign a probability weight to each microscopic configuration $\mathbf{r}$ that maps into the same CG configuration $\mathbf{R}$. This is achieved by defining the back-mapped probability

$$P_r(\mathbf{r} | U) = P_R(\mathcal{M}(\mathbf{r}) | U) \frac{g(\mathbf{r})}{\Omega(\mathcal{M}(\mathbf{r}))}, \quad (6.46)$$

where

$$\Omega(\mathbf{R}) = \int_{\Omega} d\mathbf{r} g(\mathbf{r}) \delta(\mathcal{M}(\mathbf{r}) - \mathbf{R}), \quad (6.47)$$

is the weighted degeneracy associated with the CG configuration $\mathbf{R}$. In this way, all microscopic configurations compatible with the same coarse state contribute to the lifted model distribution, with weights determined by the function $g(\mathbf{r})$. Different choices of $g(\mathbf{r})$ are possible; the most natural ones are the uniform weight $g(\mathbf{r}) = 1$, corresponding to the original relative-entropy construction (Shell, 2008), and the Boltzmann weight $g(\mathbf{r}) = e^{-\beta u(\mathbf{r})}$ (Rudzinski et al., 2011; Chaimovich et al., 2011; Noid, 2013).

In the present discussion, the second choice is the most convenient, since for $g(\mathbf{r}) \propto p_r(\mathbf{r})$, the quantity $\Omega(\mathbf{R})$ becomes proportional to the probability density induced on the reduced space by the microscopic ensemble. Recalling the definition already introduced in Chapter 2, this is precisely the mapped distribution

$$p_R(\mathbf{R}) = \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \delta(\mathcal{M}(\mathbf{r}) - \mathbf{R}). \quad (6.48)$$

With this choice, Eq. (6.46) takes the form

$$P_r(\mathbf{r} | U) = p_r(\mathbf{r}) \frac{P_R(\mathcal{M}(\mathbf{r}) | U)}{p_R(\mathcal{M}(\mathbf{r}))}. \quad (6.49)$$

The back-mapped probability is thus obtained by reweighting the microscopic distribution by the ratio between the CG model probability and the exact reduced probability associated with the microscopic ensemble.

Substituting Equation 6.49 into the definition of the relative entropy yields

$$S_{\text{rel}} = \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{p_r(\mathbf{r})}{P_r(\mathbf{r} | U)} = \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{p_R(\mathcal{M}(\mathbf{r}))}{P_R(\mathcal{M}(\mathbf{r}) | U)}. \quad (6.50)$$

Since the argument of the logarithm depends on $\mathbf{r}$ only through the reduced variable $\mathbf{R} = \mathcal{M}(\mathbf{r})$, the integral can be redefined on the CG space alone:

$$S_{\text{rel}} = \int_{\Omega_R} d\mathbf{R} p_R(\mathbf{R}) \ln \frac{p_R(\mathbf{R})}{P_R(\mathbf{R} | U)}. \quad (6.51)$$

In this form, the relative entropy is again a KL divergence, now comparing directly the exact reduced distribution induced by the microscopic system and the

one generated by the approximate CG model.

At this stage, the original variational interpretation of the method becomes clear: minimising S_{rel} with respect to $U(\mathbf{R})$ corresponds to identifying the effective CG model whose probability distribution is closest, in the KL sense, to the exact reduced distribution. This model-fitting aspect, however, is not the main object of interest here. What matters for the present thesis is that the same formalism naturally suggests how to isolate a contribution that depends on the mapping alone.

To proceed further, it is convenient to restrict the discussion to the class of *decimation mappings*. In this case, the reduced configuration $\mathbf{R}$ is obtained by retaining N of the original n degrees of freedom, while the remaining $n - N$ variables are discarded. Denoting the discarded variables collectively by $\mathbf{q}$, the microscopic configuration can then be written as $\mathbf{r} = (\mathbf{R}, \mathbf{q})$. Under this assumption, and provided that the measure factorises between retained and discarded degrees of freedom, the geometric degeneracy associated with the unresolved variables becomes independent of the specific value of $\mathbf{R}$ and reduces to a constant volume factor,

$$\Omega(\mathbf{R}) = V^{n-N}. \quad (6.52)$$

Within this restricted but practically relevant setting, the mapping-dependent contribution can be isolated in a particularly transparent form. One may then rewrite Equation 6.50 as

$$S_{\text{rel}} = \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{V^{n-N} p_r(\mathbf{r})}{p_R(\mathcal{M}(\mathbf{r}) | U)} - \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{V^{n-N} p_r(\mathbf{r})}{p_R(\mathcal{M}(\mathbf{r}))}, \quad (6.53)$$

where the factor V^{n-N} has been introduced, as in the standard derivation, to make the arguments of the logarithms dimensionless. The first term depends explicitly on the chosen CG model U , whereas the second depends only on the mapping through the induced reduced distribution $p_R(\mathbf{R})$. It is this second contribution that defines the **Mapping Entropy**:

$$S_{\text{map}} \equiv \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{V^{n-N} p_r(\mathbf{r})}{p_R(\mathcal{M}(\mathbf{r}))}. \quad (6.54)$$

By construction, S_{map} is a functional of the mapping alone. It quantifies the information loss induced by replacing the microscopic description with the reduced one, independently of any subsequent optimisation of an effective interaction model (Giulini et al., 2020).

6.4.2 Information-Theoretic Interpretation

Equation 6.54 already shows that the ME depends only on the microscopic distribution and on the reduced representation induced by the mapping. Its meaning becomes clearer once one makes explicit what information is preserved by the re-

duced description and what is instead discarded.

In the decimation setting introduced above, retaining only the variables collected in $\mathbf{R}$ means that all microscopic distinctions along the discarded coordinates are lost. Equivalently, the mapping partitions the microscopic configurational space into classes of configurations that share the same reduced representation. A natural way to formalise this loss of resolution is therefore to construct, on the original microscopic space, a distribution that preserves the probability of each **CG** state while remaining completely structureless along the unresolved coordinates. To this end, let us introduce the auxiliary distribution

$$\bar{p}_r(\mathbf{r}) \equiv \frac{1}{V^{n-N}} p_R(\mathcal{M}(\mathbf{r})). \quad (6.55)$$

This distribution is obtained by *lifting* the mapped distribution $p_R(\mathbf{R})$ back to the microscopic space and assigning it uniformly over the $n - N$ discarded degrees of freedom. In other words, $\bar{p}_r(\mathbf{r})$ retains exactly the information encoded in the reduced variable $\mathbf{R}$, while removing any residual distinction between microscopic configurations that share the same image under $\mathcal{M}$.

Using the definition of the mapped distribution (Equation 6.48), the same quantity can be written as

$$\bar{p}_r(\mathbf{r}) = \frac{1}{V^{n-N}} \int_{\Omega} d\mathbf{r}' p_r(\mathbf{r}') \delta(\mathcal{M}(\mathbf{r}') - \mathcal{M}(\mathbf{r})), \quad (6.56)$$

which makes its meaning more transparent. For a given microscopic configuration $\mathbf{r}$, the value of $\bar{p}_r(\mathbf{r})$ is obtained by collecting the total probability weight of all configurations $\mathbf{r}'$ that are mapped onto the same **CG** state as $\mathbf{r}$, and redistributing that weight uniformly within the corresponding equivalence class. The mapping therefore acts by preserving the **CG** probability mass while flattening the unresolved microscopic degrees of freedom.

With this construction, Equation 6.54 can be rewritten as

$$S_{\text{map}} = \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln \frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{r})}. \quad (6.57)$$

The **ME** is thus the **KL** divergence between the original microscopic distribution $p_r(\mathbf{r})$ and the reconstructed distribution $\bar{p}_r(\mathbf{r})$ obtained from the reduced description alone (Kullback et al., 1951; Giulini et al., 2020). In this form, its interpretation is immediate: S_{map} measures how different the true microscopic ensemble is from the best representation that can be built once only the mapped variable $\mathcal{M}(\mathbf{r})$ is kept and all finer distinctions are discarded.

As a consequence of the non-negativity of the **KL** divergence,

$$S_{\text{map}} \geq 0, \quad (6.58)$$

with equality if and only if, almost everywhere,

$$p_r(\mathbf{r}) = \bar{p}_r(\mathbf{r}). \quad (6.59)$$

This limiting case has a simple physical meaning: within each equivalence class induced by the mapping, all microscopic configurations already have the same probability density. In that situation, the mapping removes no relevant information, because the configurations that are grouped were already indistinguishable from the probabilistic point of view.

The **KL** form also clarifies the role of the mapping itself. The quantity $\bar{p}_r(\mathbf{r})$ is not an independently chosen model, but the microscopic distribution that is uniquely associated with the reduced representation once one imposes complete ignorance about the discarded degrees of freedom. The **ME** therefore does not quantify the quality of an approximate **CG** force field; rather, it quantifies the irreducible information loss due solely to the projection defined by $\mathcal{M}$.

The same quantity admits an equivalent entropic interpretation (Giulini et al., 2020; Foley et al., 2015). Let us introduce the non-ideal configurational entropy of the microscopic description,

$$s_r = - \int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \ln(V^n p_r(\mathbf{r})), \quad (6.60)$$

and the corresponding quantity for the reduced description,

$$s_R = - \int_{\Omega_R} d\mathbf{R} p_R(\mathbf{R}) \ln(V^N p_R(\mathbf{R})), \quad (6.61)$$

it can be shown (Giulini et al., 2020) that

$$S_{\text{map}} = s_R - s_r. \quad (6.62)$$

This expression shows that the **ME** can be read as the increase in configurational entropy generated by the reduction of description. Passing from the microscopic ensemble to the reduced one amounts to discarding part of the available information; correspondingly, the reduced description is compatible with a larger set of unresolved microscopic configurations and, therefore, has a larger entropy ($s_R \geq s_r$).

Equation 6.57 and Equation 6.62 provide two complementary views of the same quantity. In the first, S_{map} is an information-theoretic distance between the true microscopic ensemble and its reconstruction from the reduced description. In the second, it is the entropic cost of replacing the full-resolution description with a reduced one (Giulini et al., 2020; Foley et al., 2015). These two perspectives are fully equivalent and together make clear why the **ME** is an intrinsic property of the mapping: it depends only on how the projection groups microscopic configurations and on the probability measure carried by the original ensemble, independently of any subsequent optimisation of an effective **CG** model. For this reason, it provides a

natural criterion to compare different reduced representations solely in terms of the information they retain about the reference system.

6.4.3 Practical Evaluation of the Mapping Entropy

The exact definition of the **ME** introduced above is conceptually complete. However, its direct evaluation is generally unfeasible in realistic high-dimensional systems. In its original form, S_{map} is defined as a **KL** divergence between the microscopic distribution $p_r(\mathbf{r})$ and the corresponding back-mapped distribution $\bar{p}_r(\mathbf{r})$. This definition involves the evaluation of a logarithmic average over the full microscopic space Ω . Except for very simple cases, such a quantity cannot be computed analytically and is also numerically delicate to estimate directly. For this reason, practical applications rely on estimators computable from sampled configurations and from an explicit partition of the reduced configurational space into **CG** macrostates (Giulini et al., 2020; Giulini, 2022; Giulini et al., 2024).

A convenient starting point is the decomposition of the total **ME** into contributions associated with individual reduced configurations:

$$S_{\text{map}} = \int_{\Omega_R} d\mathbf{R} p_R(\mathbf{R}) S_{\text{map}}(\mathbf{R}), \quad (6.63)$$

where $\mathbf{R} = \mathcal{M}(\mathbf{r})$ denotes the reduced configuration induced by mapping $\mathcal{M}$, and $p_R(\mathbf{R})$ is the exact probability density on the reduced space. This representation makes it explicit that the total information loss is obtained by averaging the local contribution $S_{\text{map}}(\mathbf{R})$ over the reduced ensemble. Practical strategies for estimating the **ME** are developed from this macrostate-resolved form (Giulini et al., 2020; Giulini, 2022).

Cumulant Expansion Strategy

The first strategy applies to systems sampled at thermal equilibrium, for which the microscopic distribution is canonical:

$$p_r(\mathbf{r}) = \frac{e^{-\beta u(\mathbf{r})}}{Z}, \quad (6.64)$$

where $u(\mathbf{r})$ is the microscopic potential energy, $\beta = (k_B T)^{-1}$, and $Z = \int_{\Omega} d\mathbf{r} e^{-\beta u(\mathbf{r})}$ is the partition function.

In this setting, the **ME** associated with a given reduced configuration $\mathbf{R}$ can be expressed in terms of the conditional distribution of the microscopic energy at fixed $\mathbf{R}$ (Giulini et al., 2020; Giulini, 2022), defined as

$$P_{\beta}(U | \mathbf{R}) = \frac{\int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \delta(\mathcal{M}(\mathbf{r}) - \mathbf{R}) \delta(u(\mathbf{r}) - U)}{\int_{\Omega} d\mathbf{r} p_r(\mathbf{r}) \delta(\mathcal{M}(\mathbf{r}) - \mathbf{R})}. \quad (6.65)$$

Using this quantity, one obtains

$$S_{\text{map}}(\mathbf{R}) = k_B \ln \left[\int dU P_\beta(U | \mathbf{R}) e^{\beta(U - \langle U \rangle_{\beta|R})} \right], \quad (6.66)$$

where

$$\langle U \rangle_{\beta|R} = \int dU U P_\beta(U | \mathbf{R}), \quad (6.67)$$

is the conditional average of the microscopic potential energy restricted to the macrostate $\mathbf{R}$. The Boltzmann constant k_B is introduced to enforce consistency with the standard Gibbs formula for the entropy in statistical mechanics (Shannon, 1948).

This expression shows that, at equilibrium, the ME is controlled by the distribution of atomistic energies among the microscopic configurations that are indistinguishable at the reduced level. If all microscopic configurations mapping onto the same $\mathbf{R}$ have the same energy, then $P_\beta(U | \mathbf{R}) = \delta(U - \bar{u}_R)$ for some $\bar{u}_R$. In this case, one obtains $S_{\text{map}}(\mathbf{R}) = 0$. In this sense, zero ME is thus achieved for isoenergetic mappings (Giulini et al., 2020).

Although the previous expression is exact, its numerical evaluation is made difficult by the presence of an exponential average, whose estimation is typically affected by large sampling errors (Klimovich et al., 2015). A standard way to overcome this problem is to perform a cumulant expansion of the logarithm and truncate it at second order (Park et al., 2004; Giulini et al., 2020; Giulini, 2022). One then obtains

$$S_{\text{map}}(\mathbf{R}) \approx k_B \frac{\beta^2}{2} \left\langle (U - \langle U \rangle_{\beta|R})^2 \right\rangle_{\beta|R}, \quad (6.68)$$

which, upon averaging over the reduced distribution, leads to

$$S_{\text{map}} \approx k_B \frac{\beta^2}{2} \int_{\Omega_R} d\mathbf{R} p_R(\mathbf{R}) \left\langle (U - \langle U \rangle_{\beta|R})^2 \right\rangle_{\beta|R}. \quad (6.69)$$

The meaning of this approximation is immediate: the ME is estimated as a weighted average of the conditional energy variance within each macrostate. Accordingly, mappings that group together microscopic configurations with similar energies are associated with smaller values of S_{map} .

From a computational viewpoint, the cumulant-expansion strategy requires: (i) a set of microscopic configurations $\mathbf{r}_i$ sampled according to $p_r(\mathbf{r})$, (ii) the corresponding atomistic energies $u(\mathbf{r}_i)$, and (iii) a practical prescription to associate sampled configurations with reduced macrostates. The last point is non-trivial. The macrostates entering the formal definition are not known *a priori* and cannot, in general, be obtained by explicitly integrating over the discarded degrees of freedom. In practical implementations, one therefore constructs an empirical partition of the sampled configurations by clustering them on the basis of the retained degrees of freedom only (Giulini, 2022; Giulini et al., 2024). Once such a partition has been obtained, one estimates within each cluster the conditional mean and variance

of the atomistic energy and combines them according to the discretised counterpart of the previous equation.

It is important to stress that this route is approximate in two distinct respects. First, the exact continuous reduced space is replaced by a finite partition obtained from the available data. Second, even after such a partition has been fixed, the expression above relies on a second-order truncation of the cumulant expansion. The resulting quantity should therefore be regarded as an equilibrium estimator of the **ME**, particularly suited to situations in which equilibrium trajectories and microscopic energies are available.

Kullback-Leibler Strategy on a Discrete State Space

A second practical route is based more directly on the original information-theoretic definition of the **ME** and is formulated on a discrete set of representative microscopic configurations. This strategy is useful when one does not wish to rely on the equilibrium cumulant approximation, or more generally when the available data are already given in the form of a discrete ensemble endowed with probabilities (Giulini et al., 2024; Aldrigo et al., 2025).

Let $\{\mathbf{r}_i\}_{i=1}^L$ be a finite set of microscopic configurations, with associated probabilities $p_r(\mathbf{r}_i)$ satisfying

$$\sum_{i=1}^L p_r(\mathbf{r}_i) = 1. \quad (6.70)$$

Given a mapping $\mathcal{M}$, these configurations are grouped into **CG** macrostates by partitioning them according to the retained degrees of freedom. Let C_α denote one such macrostate. Its probability is then

$$p_R(\mathbf{R}_\alpha) = \sum_{\mathbf{r}_j \in C_\alpha} p_r(\mathbf{r}_j). \quad (6.71)$$

The discrete analogue of the back-mapped distribution is obtained by redistributing the probability of each macrostate uniformly over all microscopic configurations belonging to it. In complete analogy with the continuous construction of $\bar{p}_r(\mathbf{r})$ introduced above, one defines

$$\bar{p}_r(\mathbf{r}_i) = \frac{p_R(\mathbf{R}_\alpha)}{|C_\alpha|}, \quad \text{for } \mathbf{r}_i \in C_\alpha, \quad (6.72)$$

where $|C_\alpha|$ denotes the number of sampled microscopic configurations contained in the cluster C_α .

The **ME** estimator is then written as the discrete **KL** divergence

$$\hat{S}_{\text{map}} = k_B \sum_{i=1}^L p_r(\mathbf{r}_i) \ln \frac{p_r(\mathbf{r}_i)}{\bar{p}_r(\mathbf{r}_i)}. \quad (6.73)$$

This expression is the natural discrete counterpart of the continuous definition in Equation 6.57. It quantifies the loss of information induced by replacing, within each reduced macrostate, the original microscopic distribution by a uniform redistribution over all microscopic configurations compatible with the same reduced description.

The conceptual distinction with respect to the cumulant-expansion strategy should be emphasised. In the equilibrium route, the ME is inferred indirectly from the fluctuations of the microscopic energy inside each macrostate, exploiting the canonical form of $p_r(\mathbf{r})$. In the KL strategy, by contrast, no explicit use is made of the microscopic potential energy. What is required instead is a representative discrete ensemble together with a probability assignment on it. For this reason, the KL-based estimator can also be employed in situations where the samples are not generated by a known equilibrium dynamics, provided that the probabilities $p_r(\mathbf{r}_i)$ can be estimated or assigned.

Operationally, the procedure again relies on the construction of a partition induced by the mapping. The microscopic configurations are first projected onto the retained degrees of freedom, then grouped into clusters identified with reduced macrostates, and finally reweighted according to the average probability within each cluster. Once the discrete back-mapped distribution $\bar{p}_r(\mathbf{r}_i)$ has been constructed, the ME is evaluated by the finite sum above. This makes the KL strategy particularly convenient when the available data naturally take the form of a catalogue of representative configurations rather than a long equilibrium trajectory.

Binary-State Systems and Decimation Mappings

The KL-based formulation of the ME is particularly useful when the microscopic degrees of freedom have discrete states to begin with. Let us address the simplest of such cases, namely when the microscopic configuration is described by a set of binary variables (Aldrigo et al., 2025). Let

$$\phi = (s_1, \dots, s_n), \quad s_i \in \{-1, +1\}, \quad (6.74)$$

denote a microscopic configuration of the system, and let $p(\phi)$ be the associated probability distribution, normalised as

$$\sum_{\phi} p(\phi) = 1. \quad (6.75)$$

In this setting, a reduced representation is defined through a decimation mapping $\mathcal{M}$, which selects a subset of $N < n$ microscopic variables to be retained. If the retained indices are $i_1, \dots, i_N$, the reduced configuration reads

$$\hat{\phi} = \mathcal{M}(\phi) = (s_{i_1}, \dots, s_{i_N}). \quad (6.76)$$

The reduced variables thus identify a **CG** description obtained by discarding the remaining $n - N$ degrees of freedom. Equivalently, the mapping can be represented by a binary selection vector specifying which microscopic variables are retained and which are integrated out.

Once the mapping $\mathcal{M}$ has been fixed, it induces a probability distribution in the reduced space. Such distribution is obtained by marginalising the microscopic probability over all configurations compatible with the same reduced label:

$$P(\hat{\phi}) = \sum_{\phi} p(\phi) \delta_{\hat{\phi}, \mathcal{M}(\phi)}, \quad (6.77)$$

where $\delta_{\hat{\phi}, \mathcal{M}(\phi)}$ is equal to 1 if ϕ is mapped onto $\hat{\phi}$, and vanishes otherwise. In analogy with the general construction previously introduced, one can define the back-mapped distribution by redistributing the probability of each reduced configuration uniformly over all microscopic configurations that project onto it.

To make this construction explicit, let $\Omega(\hat{\phi})$ denote the degeneracy of the reduced label $\hat{\phi}$, namely the number of microscopic configurations that are compatible with it:

$$\Omega(\hat{\phi}) = \sum_{\phi} \delta_{\hat{\phi}, \mathcal{M}(\phi)}. \quad (6.78)$$

The corresponding back-mapped distribution is therefore

$$\bar{p}(\phi) = \frac{P(\hat{\phi})}{\Omega(\hat{\phi})}, \quad \text{with } \hat{\phi} = \mathcal{M}(\phi). \quad (6.79)$$

This expression is the discrete counterpart of the back-mapped probability density $\bar{p}_r(r)$ introduced in the general formulation: all microscopic configurations that are indistinguishable at the reduced level are assigned the same probability weight.

The **ME** associated with the mapping $\mathcal{M}$ is then written as the discrete **KL** divergence between the microscopic distribution and its back-mapped approximation:

$$S_{\text{map}}(\mathcal{M}) = k_B \sum_{\phi} p(\phi) \ln \frac{p(\phi)}{\bar{p}(\phi)}. \quad (6.80)$$

Substituting the definition of $\bar{p}(\phi)$ into the previous expression gives

$$S_{\text{map}}(\mathcal{M}) = k_B \sum_{\phi} p(\phi) \ln \frac{p(\phi) \Omega(\mathcal{M}(\phi))}{P(\mathcal{M}(\phi))}. \quad (6.81)$$

This form makes explicit that the information loss associated with a given decimation mapping is determined by two ingredients: the probability of the reduced configurations and the number of microscopic states that are merged into each of them.

It is useful to rewrite the previous expression in a form that separates the depen-

dence on the microscopic distribution from the contribution of the reduced representation. Introducing the Shannon entropy of the microscopic distribution (Shannon, 1948),

$$H_S^\phi = - \sum_{\phi} p(\phi) \ln p(\phi), \quad (6.82)$$

and the Shannon entropy of the reduced distribution,

$$H_S = - \sum_{\hat{\phi}} P(\hat{\phi}) \ln P(\hat{\phi}), \quad (6.83)$$

one obtains

$$S_{\text{map}}(\mathcal{M}) = -H_S^\phi + H_S + \sum_{\hat{\phi}} P(\hat{\phi}) \ln \Omega(\hat{\phi}). \quad (6.84)$$

This decomposition highlights the interplay between the resolution of the reduced description, measured by H_S , and the configurational degeneracy induced by the decimation procedure. In particular, at fixed number of retained variables N , variations in the ME among different mappings are determined by how the selected subset reorganises the microscopic probability mass over the reduced labels (Aldrigo et al., 2025).

In practical applications, the microscopic distribution $p(\phi)$ is typically represented by the empirical frequencies of the sampled binary configurations. For a fixed decimation mapping $\mathcal{M}$, the reduced distribution $P(\hat{\phi})$ is then obtained by grouping together all microscopic configurations that share the same retained pattern, while $\Omega(\hat{\phi})$ counts how many microscopic configurations are associated with that reduced label. The corresponding back-mapped distribution $\bar{p}(\phi)$ is finally constructed by uniformly redistributing the probability of each reduced configuration over all microscopic states compatible with it. This provides an operational route to the evaluation of the ME for discrete-state systems (Aldrigo et al., 2025).

Optimisation of the Mapping

Once the ME has been specialised to the case of discrete microscopic variables, the identification of an informative reduced description can be formulated as an optimisation problem in the space of mappings (Aldrigo et al., 2025; Giulini, 2022). At fixed number N_{cg} of retained units, one seeks the decimation mapping $\mathcal{M}$ that minimises the information loss,

$$\mathcal{M}^\star = \arg \min_{\mathcal{M} \in \mathcal{M}_{N_{\text{cg}}}} S_{\text{map}}(\mathcal{M}), \quad (6.85)$$

where $\mathcal{M}_{N_{\text{cg}}}$ denotes the set of all mappings retaining exactly N_{cg} variables out of the n available ones.

Although the discussion here focuses on a discrete setting, the optimisation of the ME applies more generally. In particular, it has also been developed for con-

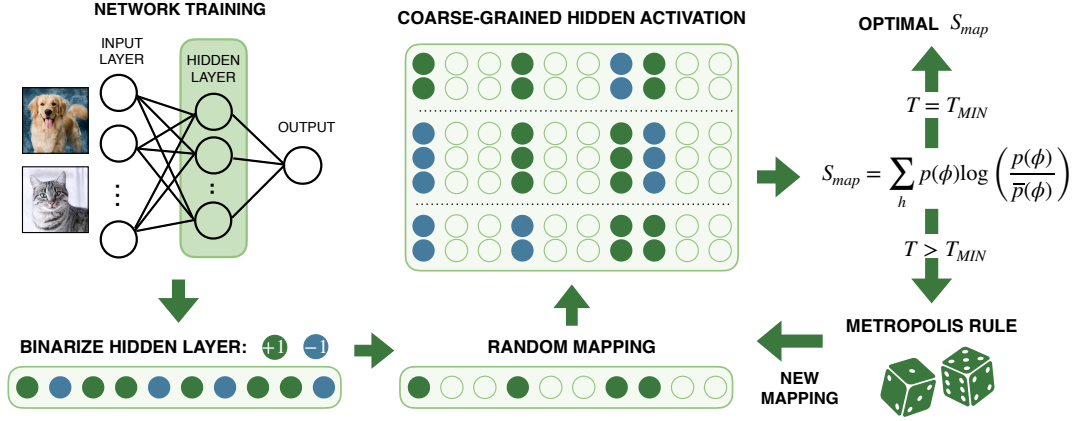


Figure 6.5: Optimisation workflow for the minimisation of the ME in binary hidden-layer representations. The input data are processed through a fully connected feed-forward neural network with one hidden layer. The hidden activations are then binarised by applying the sign function, yielding an ensemble of binary configurations of dimension n . A trial CG representation is defined by a mapping that retains $N_{cg} < n$ hidden units, thereby projecting the full hidden layer activity onto a reduced configuration. For the selected mapping, the corresponding value of S_{map} is computed from the empirical distribution of the observed hidden configurations and from the associated back-mapped distribution. The mapping is then iteratively updated through a Metropolis acceptance rule within a simulated annealing scheme, so as to explore the space of admissible subsets at fixed reduced dimensionality and identify mappings associated with low information loss.

tinuous molecular systems in the context of CG based on configurations extracted from MD trajectories (Giulini et al., 2020; Giulini et al., 2024; Giulini, 2022). In that case, for each proposed mapping, the objective function is evaluated by constructing CG macrostates through a clustering procedure on the retained continuous coordinates. In the following chapters, the same framework is instead specialised to binary hidden-layer activations, for which the reduced states are naturally defined by decimation.

In the scenario of interest here, each microscopic configuration is a binary activation pattern of the hidden layer,

$$\phi^\mu = (s_1^\mu, \dots, s_N^\mu), \quad \mu = 1, \dots, P, \quad (6.86)$$

where P is the number of sampled input patterns in the dataset, and $s_i^\mu \in \{-1, +1\}$ denotes the binarised activation of hidden unit i for sample μ , obtained by applying a sign threshold to the corresponding continuous output. A mapping $\mathcal{M}$ is therefore identified with a subset of N_{cg} hidden neurons, and the corresponding reduced configuration $\hat{\phi}^\mu = \mathcal{M}(\phi^\mu)$ is obtained by retaining only the components of ϕ^μ associated with such subset. For each choice of $\mathcal{M}$, the value of $S_{map}(\mathcal{M})$ is evaluated through the discrete KL formulation introduced above, by reconstructing the empirical reduced distribution and the corresponding degeneracy factors from the ensemble of observed binary activation patterns.

For small systems, the optimisation can in principle be carried out by exhaustively enumerating all possible mappings at fixed N_{cg} . However, the number of

admissible subsets grows combinatorially with n , so that a complete exploration rapidly becomes unfeasible as the size of the hidden layer increases. In this regime, one must resort to a stochastic exploration of the mapping space. Following the strategy already adopted in previous **ME** applications, we employ a Monte Carlo simulated annealing procedure, which is particularly suitable because the mapping space is discrete and, at fixed reduced dimensionality, is naturally explored through discrete modification moves (Kirkpatrick et al., 1983; Giulini et al., 2020; Aldrigo et al., 2025).

The simulated annealing dynamics is defined on the space of subsets of retained hidden units. Starting from an initial mapping $\mathcal{M}$, obtained for instance selecting N_{cg} neurons at random, a trial mapping $\mathcal{M}'$ is generated by replacing one retained unit with one unit that is not currently part of the subset. In this way, the cardinality of the mapping is preserved throughout the optimisation, while the algorithm performs a local exploration of the space of admissible reduced representations (Kirkpatrick et al., 1983; Giulini, 2022; Aldrigo et al., 2025). Once the move $\mathcal{M} \rightarrow \mathcal{M}'$ has been proposed, the corresponding variation of the objective function,

$$\Delta S_{\text{map}} = S_{\text{map}}(\mathcal{M}') - S_{\text{map}}(\mathcal{M}), \quad (6.87)$$

is evaluated. The trial mapping is then accepted or rejected according to a Metropolis-like criterion, using the **ME** as the cost function:

$$W(\mathcal{M} \rightarrow \mathcal{M}') = \min \left\{ 1, \exp \left[-\frac{S_{\text{map}}(\mathcal{M}') - S_{\text{map}}(\mathcal{M})}{T} \right] \right\}. \quad (6.88)$$

Therefore, moves that decrease the value of the **ME** are always accepted, while moves that increase it are accepted with a probability that depends on the effective annealing temperature T . At high temperatures, the algorithm can explore the mapping space broadly and escape local minima; as the temperature is lowered, the dynamics become increasingly biased toward mappings characterised by smaller information loss.

Following the standard simulated annealing scheme adopted in previous applications of mapping-entropy optimisation, the effective temperature is decreased according to an exponential schedule (Giulini, 2022; Giulini et al., 2024; Aldrigo et al., 2025),

$$T(i) = T_0 e^{-i/\nu}, \quad (6.89)$$

where i is the optimisation step, T_0 is the initial temperature, and ν controls the decay rate. This cooling protocol gradually drives the stochastic dynamics from an initial exploratory regime toward the identification of low-**ME** mappings. Since a single optimisation run is not guaranteed to identify the global minimum or to capture the full set of degenerate low-**ME** solutions, it is convenient to perform several independent minimisations, starting from different random initial mappings,

in order to obtain a pool of candidate low-information-loss reduced representations (Giulini, 2022; Aldrigo et al., 2025). The overall workflow employed in the present discrete case is schematically illustrated in Figure 6.5.

The formal setting introduced above, together with the Wang-Landau algorithm and the Mapping Entropy framework, provides the methodological toolkit for the analyses presented in the following chapters. These tools will be used to investigate the organisation of neural network configuration spaces and the information retained by reduced hidden representations. Before moving to more direct applications, however, the next chapter is devoted to a methodological decomposition of the ME, with the aim of clarifying its structure and interpretation independently of the specific system under consideration.

7

Mapping Entropy Multi-Body Decomposition

7.1 Introduction

In the previous chapter, **Mapping Entropy (ME)** was introduced as an information-theoretic measure of the loss of information that results from replacing a detailed description of a system with a simpler one. In this form, the **ME** provides a natural criterion to compare different reduced representations and to identify those retaining the largest amount of information about the reference ensemble. This capability makes **ME** particularly relevant in the context of the present thesis, where reduced descriptions are used as a tool to probe the structure of complex systems and, in particular, to analyse the informational content of neural networks.

At the same time, however, **ME** is an intrinsically multi-body functional of the mapping, which depends on all the retained system elements at once. In quantifying the overall information loss associated with a mapping, it does not immediately reveal how this loss is distributed among the degrees of freedom, nor how it is shaped by their statistical dependencies of increasing order. As a consequence, although the **ME** can be used directly as an optimisation functional, the structure it reveals remains partially hidden. A more explicit decomposition would therefore be useful both to clarify the meaning of the quantity itself and to understand which features of the underlying system are most relevant in determining the quality of a reduced representation.

The aim of this chapter is to derive such a decomposition in the case of discrete degrees of freedom. The interest of this construction is twofold: on the one hand, it provides a more transparent interpretation of the **ME**, by relating it to more familiar information-theoretic quantities and by disentangling the contributions arising from one-body, two-body, and higher-order statistical structure (Killian et al., 2007; Matsuda, 2000). On the other hand, it suggests a natural route toward low-order

approximations of the full quantity, which may be of practical interest whenever the exact evaluation or optimisation of the **ME** becomes computationally demanding.

The chapter is organised as follows. After introducing the formal setting used for the decomposition, the multi-body expansion of the **ME** is derived and its first terms are discussed explicitly. The consequences of this result are then analysed, with particular attention to the interpretation of the lowest-order contributions and to the insight that they provide into the informational organisation of the system.

7.2 Formal Framework

The definitions and notation for the **ME** introduced in the previous chapter are adopted here in the specific case of systems with discrete degrees of freedom. In particular, the quantity is understood in the discrete **KL**-based formulation discussed there, to which the reader is referred for a more detailed treatment of the reduced distribution, the associated back-mapped distribution, and their relation to the information loss induced by a mapping.

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote a microscopic configuration of a system with n discrete degrees of freedom, distributed according to a probability law $p(\mathbf{x})$ satisfying

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1. \quad (7.1)$$

A reduced representation is obtained by retaining only a subset of the original variables. In the following, such a mapping is encoded by a binary selector vector

$$\mathbf{c} = (c_1, \dots, c_n), \quad c_i \in \{0, 1\}, \quad (7.2)$$

where $c_i = 1$ if the variable x_i is retained and $c_i = 0$ otherwise. At fixed reduced dimensionality,

$$N = \sum_{i=1}^n c_i, \quad (7.3)$$

different reduced descriptions correspond to different choices of the vector $\mathbf{c}$.

Within this setting, the **ME** can be regarded as a function of the mapping, and therefore as a function of $\mathbf{c}$. The aim of this chapter is to make this dependence explicit by deriving a decomposition in terms of contributions of increasing order in the retained variables. Formally, one seeks an expansion of the form

$$S_{\text{map}}(\mathbf{c}) = S^{(0)} + \sum_i S_i^{(1)} c_i + \sum_{i < j} S_{ij}^{(2)} c_i c_j + \sum_{i < j < k} S_{ijk}^{(3)} c_i c_j c_k + \dots, \quad (7.4)$$

where $S^{(0)}$ is a constant term, while the coefficients $S_i^{(1)}, S_{ij}^{(2)}, S_{ijk}^{(3)}, \dots$, quantify the one-body, two-body, three-body, and higher-order contributions to the information loss associated with the mapping.

No specific dynamical model is assumed in the derivation developed below. The discussion is carried out at the level of a generic discrete probability distribution and is therefore intended to clarify the structural properties of the ME itself, independently of the particular physical system to which the quantity may later be applied.

7.3 Results

In this section, the expansion of the ME in powers of the selector variables c_i is derived to make explicit the contributions associated with statistical dependencies of increasing order among the microscopic degrees of freedom. The derivation is conveniently organised in two steps. First, the entropy of a generic discrete system is rewritten in a multi-body form. This representation is then specialised to the reduced distribution entering the definition of the ME (Killian et al., 2007; Watanabe, 1960).

7.3.1 Multi-body Decomposition of the Entropy

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote a microscopic configuration distributed according to $p(\mathbf{x})$. The entropy reads

$$S = - \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln p(\mathbf{x}). \quad (7.5)$$

The probability of a configuration can be written as

$$p(\mathbf{x}) = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \mathbb{I}(\mathbf{x}, \mathbf{x}'), \quad (7.6)$$

where $\mathbb{I}(\mathbf{x}, \mathbf{x}')$ is a characteristic function equal to unity if $\mathbf{x} = \mathbf{x}'$ and vanishes otherwise. Since the identity of two configurations is equivalent to the identity of all their components, one has

$$\mathbb{I}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^n \mathbb{I}(x_i, x'_i) = \prod_{i=1}^n (1 + \delta_i), \quad (7.7)$$

where $\delta_i \equiv \mathbb{I}(x_i, x'_i) - 1$. This representation makes it possible to express the probability of a configuration as a sum of contributions of increasing order,

$$p(\mathbf{x}) = 1 + \sum_i \langle \delta_i \rangle + \sum_{i < j} \langle \delta_i \delta_j \rangle + \sum_{i < j < k} \langle \delta_i \delta_j \delta_k \rangle + \dots, \quad (7.8)$$

where the averages are taken with respect to the distribution $p(\mathbf{x}')$, namely

$$\langle f \rangle \equiv \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') f(\mathbf{x}, \mathbf{x}'). \quad (7.9)$$

For notational convenience, let $X(\mathbf{x}) \equiv \sum_i \langle \delta_i \rangle + \sum_{i < j} \langle \delta_i \delta_j \rangle + \dots$, so that Equation 7.8 can be written as

$$p(\mathbf{x}) = 1 + X(\mathbf{x}). \quad (7.10)$$

The advantage of this formulation is that it allows us to expand the logarithm entering Equation 7.5 around unity and thereby reorganise the entropy into contributions of increasing order. Indeed, using

$$\ln(1 + X) = X - \frac{1}{2}X^2 + \frac{1}{3}X^3 + \dots, \quad (7.11)$$

and substituting the resulting expression into Equation 7.5, one can collect all terms according to the number of indices involved. The zeroth-order contribution vanishes as a consequence of probability normalisation, so that the entropy takes the hierarchical form

$$-S = \sum_i Q_i + \sum_{i < j} Q_{ij} + \sum_{i < j < k} Q_{ijk} + \dots, \quad (7.12)$$

where the quantities Q_i , Q_{ij} , Q_{ijk} , and the higher-order terms are obtained by collecting all contributions involving respectively one index, two indices, three indices, and progressively larger subsets of variables. The entropy is thus reorganised into a hierarchy of terms associated with statistical structures of increasing order. The explicit derivation of these contributions is reported in Appendix E; for the purposes of the present discussion, it is sufficient to report explicitly the lowest-order terms, which already illustrate the structure of the expansion.

The first non-vanishing contribution is the one-body term, which depends only on the marginal distribution of the variable x_i ,

$$Q_i = \sum_{\{x_i\}} p(x_i) \ln p(x_i). \quad (7.13)$$

This quantity is therefore the negative entropy of the corresponding one-variable marginal. At the second order one obtains

$$Q_{ij} = \sum_{\{x_i, x_j\}} p(x_i, x_j) \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \quad (7.14)$$

which vanishes when x_i and x_j are statistically independent. The two-body term thus represents the correction associated with pairwise dependence and can be identified with the mutual information between the two variables (Cover et al., 2006). At the third order, one finds a contribution of the form

$$Q_{ijk} = \sum_{\{x_i, x_j, x_k\}} p(x_i, x_j, x_k) \ln \frac{p(x_i, x_j, x_k) p(x_i) p(x_j) p(x_k)}{p(x_i, x_j) p(x_i, x_k) p(x_j, x_k)}, \quad (7.15)$$

which accounts for the part of the statistical dependence among x_i , x_j , and x_k that

is not exhausted by the one-body and pairwise terms alone. In this sense, Q_{ijk} represents the first genuinely higher-order correction in the expansion (McGill, 1954; Matsuda, 2000). Higher-order contributions follow analogously and describe irreducible dependencies involving progressively larger subsets of degrees of freedom.

7.3.2 Multi-body Decomposition of the Mapping Entropy

The general multi-body expansion derived in the previous subsection can now be specialised to the ME. In the discrete formulation introduced in Chapter 6, this quantity is defined as

$$S_{\text{map}}(\mathbf{c}) = \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\bar{p}_{\mathbf{c}}(\mathbf{x})} \quad (7.16)$$

$$= \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln p(\mathbf{x}) - \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln q_{\mathbf{c}}(\mathbf{x}) + \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln \Omega(\mathbf{x}). \quad (7.17)$$

The second line follows by substituting the definition of the back-mapped distribution,

$$\bar{p}_{\mathbf{c}}(\mathbf{x}) = \frac{q_{\mathbf{c}}(\mathbf{x})}{\Omega(\mathbf{x})}, \quad (7.18)$$

where $q_{\mathbf{c}}(\mathbf{x})$ denotes the probability of the CG configuration selected by the mapping, and $\Omega(\mathbf{x})$ its degeneracy, namely the number of microscopic configurations compatible with the same reduced state. In the explicit form given in Equation 7.17, the ME is thus naturally separated into three contributions. The first term is the entropy of the full distribution, whose multi-body decomposition has already been derived in the previous subsection. The remaining task is therefore to specialise the same treatment to the reduced probability $q_{\mathbf{c}}(\mathbf{x})$, where the dependence on the selector variables appears explicitly, and then combine the resulting expansion with the contribution arising from the degeneracy term $\Omega(\mathbf{x})$.

The first step is, therefore, to treat the term depending on the reduced probability $q_{\mathbf{c}}(\mathbf{x})$. This quantity is obtained by summing over all microscopic configurations compatible with the CG state selected by the mapping, and can be written in the factorised form

$$q_{\mathbf{c}}(\mathbf{x}) = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \prod_{i=1}^n [1 + c_i \delta_i], \quad (7.19)$$

where δ_i encodes the matching condition between the configurations $\mathbf{x}$ and $\mathbf{x}'$ at site i . The selector variable c_i specifies whether site i is retained by the mapping: when $c_i = 1$, the corresponding matching condition is included in the product, whereas for $c_i = 0$ it is omitted.

Written in this form, $q_{\mathbf{c}}(\mathbf{x})$ has the same algebraic structure as the full probability $p(\mathbf{x})$ leading to the decomposition in Equation 7.8. The derivation of the logarithmic expansion therefore carries over unchanged, with the selector variables appearing

only as multiplicative weights. As a result,

$$\sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln q_{\mathbf{c}}(\mathbf{x}) = \sum_i Q_i c_i + \sum_{i < j} Q_{ij} c_i c_j + \sum_{i < j < k} Q_{ijk} c_i c_j c_k + \dots \quad (7.20)$$

The coefficients Q_i , Q_{ij} , Q_{ijk} , ... are given by the same expressions already derived in Equation 7.13-7.14-7.15, the dependence on the mapping entering only through the selector products.

The last contribution in Equation 7.17 comes from the degeneracy factor $\Omega(\mathbf{x})$. For a decimation mapping, this quantity counts the number of microscopic configurations that are compatible with the same reduced configuration, namely those sharing the same values of the retained variables. At this point, we assume that the accessible configurational space factorises as the direct product of the local state spaces of the individual degrees of freedom. Thus, if each degree of freedom x_i can access ω_i distinct states, every combination of such local states is in principle an admissible microscopic configuration, and the total size of the configurational space is

$$\Omega_0 = \prod_{i=1}^n \omega_i. \quad (7.21)$$

This assumption is exact for systems whose admissible configurations are fully specified by independent local state spaces, and it is often a natural starting point in discrete model systems. In empirical settings, however, the actually visited configuration space may be only a subset of this full product space, either because some combinations are excluded by correlations or constraints, or because finite sampling does not explore all admissible states.

For a given mapping $\mathbf{c}$, only the variables that are not retained contribute to the degeneracy, so that

$$\Omega(\mathbf{x}) = \prod_{i=1}^n \omega_i^{1-c_i}. \quad (7.22)$$

Indeed, when $c_i = 1$ the value of x_i is fixed by the mapping and no degeneracy is associated with that degree of freedom, whereas for $c_i = 0$ all its ω_i accessible states contribute.

The corresponding term in the ME can then be evaluated directly

$$\sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln \Omega(\mathbf{x}) = S_0 - \sum_{i=1}^n c_i \ln \omega_i, \quad (7.23)$$

where

$$S_0 = \ln \Omega_0 = \sum_{i=1}^n \ln \omega_i, \quad (7.24)$$

is the entropy of the corresponding uniform reference system, namely the system

defined on the same (factorised) configurational space of size Ω_0 , in which all microscopic configurations are assigned equal probability.

By combining the expansions of the three contributions in Equation 7.17, the ME can finally be expressed as a many-body series in the selector variables,

$$S_{\text{map}}(\mathbf{c}) = S_0 - S - \sum_i (Q_i + \ln \omega_i) c_i - \sum_{i < j} Q_{ij} c_i c_j - \sum_{i < j < k} Q_{ijk} c_i c_j c_k - \dots \quad (7.25)$$

Equation 7.25 makes explicit how the ME depends on the subset of retained variables. In particular, two limiting cases are recovered immediately. When $c_i = 0$ for all i , no variable is retained and all selector-dependent contributions vanish. This corresponds to the maximal ME, since the reduced distribution becomes uniform over all microscopic configurations compatible with the mapping and therefore carries no information about the original state. In this limit, the ME reduces to the negative excess entropy of the full system, namely to the difference between the entropy of the corresponding uniform reference distribution and that of the microscopic distribution,

$$S_{\text{map}}(\mathbf{0}) = S_0 - S. \quad (7.26)$$

At the opposite extreme, when $c_i = 1$ for all i , all degrees of freedom are retained. Using the decomposition of the full entropy S derived in the previous subsection together with $S_0 = \sum_i \ln \omega_i$, is easy to demonstrate that S_{map} vanishes identically, as expected for the fully resolved mapping.

Between these two extremes, Equation 7.25 provides a direct interpretation of the information progressively recovered by the mapping. The one-body contributions quantify the reduction of ME associated with retaining individual variables, while the two-body terms account for the additional information recovered when pairwise statistical dependencies are resolved. Both contributions are necessarily negative: $(Q_i + \ln \omega_i)$ is a KL divergence (see Appendix E) and Q_{ij} a mutual information, and are therefore positive semidefinite. Their effect is therefore to lower the ME from its maximal value as more structure of the original distribution is retained.

The interpretation becomes qualitatively different starting from the third order. The coefficient Q_{ijk} is not sign-definite and may be either positive or negative (McGill, 1954; Matsuda, 2000). Depending on the statistical structure of the system, they may either reinforce the information recovered at lower orders or partially compensate it.

7.3.3 Non-Interacting Model

To better illustrate the meaning of the decomposition derived above, it is useful to consider a simple reference case in which the microscopic degrees of freedom are statistically independent. Despite its apparent simplicity, this model is instructive

because the corresponding ME can be computed exactly, thereby clarifying several non-trivial properties of the quantity itself.

Consider a system of n independent binary spins, $\sigma_i \in \{0, 1\}$. For each spin, let $p_i \equiv p(\sigma_i = 1)$, so that the corresponding marginal distribution is

$$p(\sigma_i) = p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i}. \quad (7.27)$$

Statistical independence then implies that the probability of a microscopic configuration $\sigma \equiv (\sigma_1, \dots, \sigma_n)$ factorises as

$$p(\sigma) = \prod_{i=1}^n p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i}. \quad (7.28)$$

The same property carries over to the reduced probability. Indeed, substituting Equation 7.28 into the general expression Equation 7.19 gives

$$q(\sigma) = \prod_{i=1}^n [1 + c_i (p_i^{\sigma_i} (1 - p_i)^{1 - \sigma_i} - 1)], \quad (7.29)$$

so that both $p(\sigma)$ and $q(\sigma)$ consist of products of single-spin contributions.

According to the general definition of the ME in Equation 7.17, the calculation is therefore reduced to evaluating three terms: the entropy of the full distribution, the contribution involving the reduced probability $q(\sigma)$, and the degeneracy term.

The first two contributions in Equation 7.17 can be treated together. Because both $p(\sigma)$ and $q(\sigma)$ factorise, they give rise to the same single-spin structure, the only difference being the presence of the selector variables c_i in the reduced term

$$\text{full} \quad \sum_{\{\sigma\}} p(\sigma) \ln p(\sigma) = \sum_{i=1}^n [p_i \ln p_i + (1 - p_i) \ln(1 - p_i)], \quad (7.30)$$

$$\text{reduced} \quad \sum_{\{\sigma\}} p(\sigma) \ln q(\sigma) = \sum_{i=1}^n c_i [p_i \ln p_i + (1 - p_i) \ln(1 - p_i)]. \quad (7.31)$$

The role of the selector variables is now clear. In Equation 7.30 every spin contributes, while in Equation 7.31 only the retained spins do. Their difference, therefore, isolates the contribution of the discarded degrees of freedom:

$$\sum_{\{\sigma\}} p(\sigma) \ln p(\sigma) - \sum_{\{\sigma\}} p(\sigma) \ln q(\sigma) = \sum_{i=1}^n (1 - c_i) [p_i \ln p_i + (1 - p_i) \ln(1 - p_i)]. \quad (7.32)$$

The last contribution in Equation 7.17 is the degeneracy term. In the present case, its evaluation is particularly simple. Since each discarded spin can take two possible values, every unresolved degree of freedom contributes a factor of 2 to

the number of microscopic configurations compatible with the CG state. The total degeneracy is therefore determined only by the number of discarded spins and not by the specific values of the probabilities p_i . Denoting by $N = \sum_i c_i$ the number of retained spins, it follows that $\Omega(\sigma) = 2^{n-N}$, and hence

$$\sum_{\{\sigma\}} p(\sigma) \ln \Omega(\sigma) = (n - N) \ln 2. \quad (7.33)$$

Collecting these results, the ME of the non-interacting spin system becomes

$$S_{\text{map}} = (n - N) \ln 2 - \sum_{i=1}^n (1 - c_i) [-p_i \ln p_i - (1 - p_i) \ln(1 - p_i)]. \quad (7.34)$$

Equation 7.34 provides a particularly clear interpretation of the non-interacting case. In the absence of correlations, the ME reduces to a sum of independent one-body contributions associated with the discarded spins. No pair or higher-order correction appears, consistently with the fact that the microscopic distribution contains no statistical dependencies beyond the single-site level. In this sense, the information loss induced by the mapping is entirely controlled by the local uncertainty of the unresolved degrees of freedom.

The additive form of Equation 7.34 makes the optimisation problem fully tractable. In particular, the minimum ME can be studied from two complementary perspectives: as a function of the number of retained sites N , and as a function of the resolution, introduced in Chapter 2, which provides a more physical measure of the retained description. These two viewpoints are discussed separately in the following.

Lowest Mapping Entropy as a Function of the Number of Retained Sites

The optimisation problem becomes particularly simple when the number of retained sites N is fixed. Indeed, as follows from Equation 7.34, the ME is the sum of two contributions. The first term, $(n - N) \ln 2$, depends only on *how many* sites are retained and is therefore constant at fixed N . As a result, the optimisation over *which* sites should be retained reduces to minimising the second contribution. In particular, the minimum is obtained by discarding the $n - N$ spins with the largest one-body entropy, defined as

$$s(p_i) \equiv -p_i \ln p_i - (1 - p_i) \ln(1 - p_i). \quad (7.35)$$

The function $s(p_i)$ is minimal when $p_i = 0$ or $p_i = 1$, for which $s(p_i) = 0$. In both cases, the spin is completely deterministic, since it always takes the same value and therefore carries no uncertainty. By contrast, $s(p_i)$ is maximal for $p_i = 1/2$, where $s(1/2) = \ln 2$. This corresponds to the situation of maximal uncertainty, in which the two spin states are equally probable. More generally, $s(p_i)$ decreases as p_i

moves away from $1/2$ toward either 0 or 1, reflecting the fact that the spin becomes progressively more biased and therefore more predictable. The full optimisation problem is therefore solved by ranking the sites according to their local entropy.

Denoting by $S_{\text{map}}^*(N)$ the lowest ME value obtained when retaining N sites, one can write

$$S_{\text{map}}^*(N) = (n - N) \ln 2 - \sum_{j=1}^{n-N} s(p_{i_j}), \quad (7.36)$$

where the indices i_j label the $n - N$ discarded spins ranked in order of decreasing one-body entropy, that is, their probabilities are the closest to the maximally uncertain value $p_i = 1/2$.

The behaviour of $S_{\text{map}}^*(N)$ as a function of the number of retained sites can now be analysed explicitly. When passing from N to $N+1$, one additional site is retained; in the optimal construction, this is the site with the smallest local entropy among those that were previously discarded. The difference between the two optimal values is therefore

$$\Delta S_{\text{map}}^*(N) \equiv S_{\text{map}}^*(N+1) - S_{\text{map}}^*(N) = s(p_{i^*}) - \ln 2, \quad (7.37)$$

where i^* denotes the site by which the two optimal selections differ. Since $0 \leq s(p_i) \leq \ln 2$ for any binary spin, it follows immediately that

$$\Delta S_{\text{map}}^*(N) \leq 0. \quad (7.38)$$

Hence, for the ideal non-interacting spin system, the curve of the minimum ME is monotonically non-increasing as a function of N . Physically, this is the expected behaviour: retaining one more degree of freedom can only reduce, or at most leave unchanged, the information loss associated with the mapping.

A special case occurs when one of the spins has $p_i = 1/2$. In that case its local entropy is exactly $\ln 2$, and if this spin is the one added when going from N to $N+1$, then $\Delta S_{\text{map}}^*(N) = 0$. More generally, if several spins satisfy $p_i = 1/2$, they are all equivalent from the point of view of the optimisation, and the curve $S_{\text{map}}^*(N)$ develops a flat region. Such a plateau is therefore a trivial consequence of the presence of maximally uncertain degrees of freedom, all contributing equally to the information loss, rather than a signature of collective effects.

Finite-Sampling Effects and Sampled Configurational Space

The discussion above refers to the ideal setting in which the accessible configurational space is fully represented. It is therefore crucial to understand how this picture is modified when the probability distribution is estimated from a finite sample. Even for a system that is non-interacting in principle, finite sampling may prevent a complete exploration of the configurational space. In that case, while the one-body probabilities p_i may still be estimated with reasonable accuracy, the degeneracy

term entering the **ME** need no longer attain its ideal value $(n - N) \ln 2$.

To make this point explicit, let M_N denote a mapping retaining N sites, and define

$$\hat{\Omega}_N(\sigma) = \sum_{\{\sigma'\}} \prod_{i \in M_N} \mathbb{I}(\sigma_i, \sigma'_i), \quad (7.39)$$

where the hat indicates that the sum runs only over the distinct configurations that are actually present in the sample. The quantity $\hat{\Omega}_N(\sigma)$ is thus the number of sampled microscopic configurations that map onto the same **CG** configuration as σ .

Now consider a mapping M_{N+1} obtained from M_N by retaining one additional spin j . One then has

$$\hat{\Omega}_N(\sigma) = \hat{\Omega}_{N+1}(\sigma; \sigma_j = 0) + \hat{\Omega}_{N+1}(\sigma; \sigma_j = 1). \quad (7.40)$$

If both values of the additional spin are represented in the sample for every fixed configuration of the other retained variables, then $\hat{\Omega}_N(\sigma) = 2\hat{\Omega}_{N+1}(\sigma)$. Otherwise, if some configurations are missing, one has instead $\hat{\Omega}_N(\sigma) < 2\hat{\Omega}_{N+1}(\sigma)$. Therefore,

$$\langle \ln \hat{\Omega}_{N+1}(\sigma) \rangle - \langle \ln \hat{\Omega}_N(\sigma) \rangle \geq -\ln 2, \quad (7.41)$$

with equality only in the fully sampled case.

This shows that, in the finite-sampling regime, the decrease of the minimum **ME** as N increases can be weaker than in the ideal case. As sampling becomes poorer and the configurational space is explored less completely, the negative contribution associated with the degeneracy term is reduced in magnitude. As a consequence, the curve $S_{\text{map}}^*(N)$ may develop extended nearly flat regions even in the absence of genuine correlations and even when no spin has exactly $p_i = 1/2$ (Aldrigo et al., 2025). In extreme undersampling conditions, one cannot even exclude pathological situations in which the estimated minimum **ME** increases when an additional site is retained. Such behaviour, however, should be interpreted as a property of the finite sample rather than as a structural property of the underlying non-interacting system.

Lowest Mapping Entropy as a Function of the Resolution

A complementary viewpoint is obtained by analysing the minimum **ME** at fixed number of retained sites N as a function of the *resolution* of the retained description. Within the **Resolution-Relevance** framework introduced in Chapter 2, the resolution quantifies the entropy of the **CG** representation and therefore measures how detailed, or informative, the retained description is. This representation is especially useful because, unlike N , the resolution does not only count how many degrees of freedom are kept, but also reflects how much information is effectively encoded in the corresponding **CG** variables (Marsili et al., 2013).

In general, the **ME** can be written as

$$S_{\text{map}} = H_s - S + \langle \ln \Omega_N(\sigma) \rangle, \quad (7.42)$$

where S is the entropy of the full system, while

$$H_s = - \sum_{\{\sigma\}} p(\sigma) \ln q(\sigma) = - \sum_{\{s\}} q(s) \ln q(s), \quad (7.43)$$

is the resolution, namely the entropy of the system represented in terms of the retained variables only. Here s denotes a **CG** configuration,

$$\{s\} \equiv \{\sigma_i : i \in M_N\}, \quad (7.44)$$

with M_N indicating the mapping retaining N sites.

Equation 7.42 separates the **ME** into three conceptually distinct contributions: the full-system entropy S , which is independent of the mapping; the resolution H_s , which measures the informational richness of the retained description; and the average logarithmic degeneracy $\langle \ln \Omega_N(\sigma) \rangle$, which quantifies how many microscopic configurations remain unresolved once the mapping has been applied.

In the ideal oversampling limit, where the configurational space is explored so thoroughly that all microscopic configurations compatible with a given **CG** state are effectively sampled, the degeneracy term attains its maximal value

$$\langle \ln \Omega_N(\sigma) \rangle = (n - N) \ln 2. \quad (7.45)$$

At fixed N , this is a constant independent of the specific mapping. In that regime, one therefore obtains

$$S_{\text{map}} = H_s - S + (n - N) \ln 2, \quad (7.46)$$

that is, for fixed N , the **ME** is a linear function of the resolution with unit slope. It follows immediately that, in the fully sampled limit, minimising the **ME** is equivalent to minimising the resolution. In other words, among all mappings retaining the same number of sites, the optimal one is the one whose **CG** distribution has the lowest entropy.

This behaviour has a clear physical interpretation. If all **CG** degeneracies are perfectly resolved by the data, then the only mapping-dependent contribution left is the entropy of the reduced representation itself. The lower the resolution, the more concentrated the **CG** distribution, and the smaller the information loss associated with the mapping.

The situation changes when the probability distribution is estimated from a fi-

nite sample. In that case, the empirical degeneracy $\hat{\Omega}_N(\sigma)$ generally satisfies

$$\hat{\Omega}_N(\sigma) \leq \Omega_N(\sigma) = 2^{n-N}, \quad (7.47)$$

because not all microscopic configurations compatible with a given CG state happen to be visited in the sample. The empirical counterpart of Equation 7.42 becomes

$$\hat{S}_{\text{map}} = \hat{H}_s - S + \sum_s \hat{q}(s) \ln \hat{\Omega}_N(s), \quad (7.48)$$

where

$$\hat{H}_s = - \sum_{\{s\}} \hat{q}(s) \ln \hat{q}(s), \quad \hat{q}(s) = \frac{k_s}{\mathcal{M}}, \quad (7.49)$$

$\mathcal{M}$ is the number of sampled microscopic configurations, and k_s is the number of sampled configurations that map onto the CG state s .

The important point is that finite sampling affects $\hat{q}(s)$ and $\hat{\Omega}_N(s)$ differently. The empirical CG probabilities may already be estimated with reasonable accuracy even when sampling is still insufficient to observe, for each state s , the full set of compatible microscopic configurations. Hence, $\hat{\Omega}_N(s)$ can be appreciably smaller than its ideal value.

This mismatch is expected to be more severe at high empirical resolution. Indeed, a larger $\hat{H}_s$ corresponds to a broader CG distribution, so that the available sample is spread over many states s . For fixed $\mathcal{M}$, each state is then visited fewer times on average, and its microscopic multiplicity is less completely explored. By contrast, when $\hat{H}_s$ is small, the probability is concentrated on fewer CG states. These dominant states collect more samples, so their microscopic degeneracies are estimated more reliably, while the poorly sampled rare states are suppressed by their small statistical weight.

Therefore, higher empirical resolution generally increases the impact of under-sampling on the degeneracy term, although this should be regarded as a qualitative trend rather than a strict monotonic relation. In this sense, the quantity $\ln \hat{\Omega}_N(s)$ is expected to decrease as the empirical resolution of the mapping increases.

This observation has a direct consequence for the shape of the minimum ME as a function of the resolution at fixed N . In the well-sampled regime, the degeneracy term is nearly constant, and $\hat{S}_{\text{map}}$ remains approximately linear and increasing in $\hat{H}_s$. Hence, the minimum is attained at the lowest available resolution. For sufficiently poor sampling, however, the decrease of the degeneracy term with $\hat{H}_s$ can compensate for, or even overcome, the explicit linear increase coming from the first term in Equation 7.48.

Three qualitatively distinct regimes may therefore be identified. For relatively small N , the mapping still groups many microscopic configurations into each CG state, so the system remains close to the oversampling regime, and the minimum ME is reached at low resolution. For sufficiently large N , on the contrary, the CG

representation becomes itself high-dimensional and sparse, so that undersampling becomes severe. In that case, the degeneracy term may dominate, and the minimum ME may shift toward mappings with larger resolution. Between these two limits, an intermediate regime can occur in which the decrease of the degeneracy term approximately balances the explicit increase in $\hat{H}_s$, producing an extended, nearly flat portion in the curve of the ME as a function of the resolution.

This plateau has a meaning different from the one discussed in the previous subsection. There, a flat region originated from the presence of degrees of freedom with maximal one-body entropy, which are genuinely equivalent from the point of view of the optimisation. Here, instead, a flat dependence on the resolution is not a structural property of the underlying system, but the signature of a crossover between a regime in which the CG degeneracy is reliably sampled and one in which it is not. In this sense, extended nearly flat regions have already been observed also in empirical and more complex settings, where the effective degrees of freedom are mutually correlated (Aldrigo et al., 2025). The shape of the curve is therefore controlled not only by the statistics of the system but also by the quality of sampling.

7.3.4 Computational Illustration of the Non-Interacting System

The objective of this empirical analysis is to assess whether the closed-form expression for the minimum ME, $S_{\text{map}}^*(N)$, provides an accurate estimate of the ME computed from data generated by an *independent* spin model. To this end, samples of dimension $n = 25$ are repeatedly drawn from the reference distribution specified by the single-site probabilities $\{p_i\}_{i=1}^n$, thereby constructing data sets containing P configurations. To distinguish purely statistical fluctuations from systematic finite-sampling effects, three data sets of increasing size are considered, with $P \in \{10^3, 10^5, 10^7\}$.

To evaluate the ME on these generated data sets, one must estimate the microscopic probability associated with each configuration, $p(\sigma)$ and, for any chosen set of retained degrees of freedom, it is also necessary to estimate the probability of the corresponding CG configurations, $q(s)$. Both quantities are computed exclusively from the finite ensemble of sampled configurations, following the standard procedure introduced in Chapter 6. More precisely, $p(\sigma)$ is estimated by counting the number of occurrences of configuration σ in the data set and normalising by the total sample size. Likewise, $q(s)$ is obtained by counting the occurrences of the reduced configuration s , defined by projecting each sampled configuration onto the $N < n$ retained spins.

At this stage, the CG probability $q(s)$ must be lifted back to the microscopic configuration space, since the definition of the ME involves the back-mapped distribution $\bar{p}(\sigma)$ (Equation 7.16). In general, all microscopic configurations σ that project

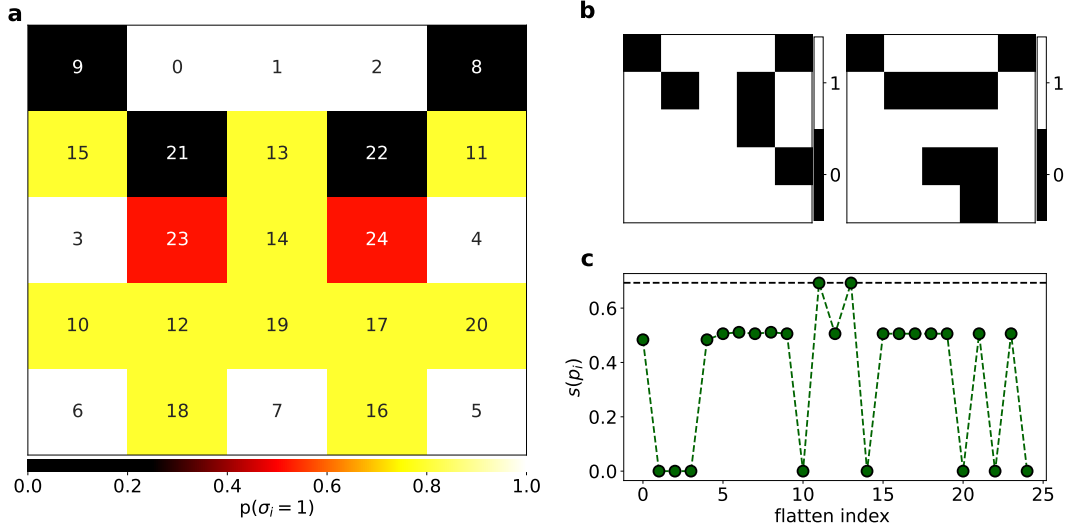


Figure 7.1: (a) Reference single-site probabilities $p(\sigma_i = 1)$ obtained from the image intensity map (colour scale). The integer shown on each pixel indicates the order in which that pixel is selected as predicted by the theoretical optimisation at fixed N (0 is the first pixel to be retained, then 1, etc.). (b) Representative configurations sampled from the corresponding independent-spin model (black/white denote $\sigma_i = 1/0$). (c) One-body entropies $s(p_i)$ associated with the reference probabilities in panel (a), shown as a function of the flattened pixel index; the horizontal dashed line marks the maximum value $s(1/2) = \ln 2$.

onto the same CG label s are assigned the same probability, so that

$$\bar{p}(\sigma) = \frac{q(s)}{\Omega(s)}, \quad (7.50)$$

where $\Omega(s)$ denotes the degeneracy, *i.e.*, the number of microscopic configurations mapping onto s . In the ideal setting, for a binary system, a mapping that retains N spins out of n , this degeneracy is $\Omega(s) = 2^{n-N}$, independently of s . However, when $\bar{p}(\sigma)$ is reconstructed from a finite sample, $\Omega(s)$ can be treated in two distinct ways, leading to the two empirical estimators compared in the following.

In the first case, the degeneracy is taken to be the *empirical* one, $\Omega_P(s)$, defined as the number of distinct microscopic configurations observed in the dataset that project onto the same CG state s . The weight $q(s)$ is then redistributed uniformly among those observed pre-images, so that $\bar{p}(\sigma) = q(s)/\Omega_P(s)$ for every sampled σ compatible with s . The corresponding ME estimate is denoted by S_{map} .

In the second case, the degeneracy is fixed to its *ideal* value 2^{n-N} , *i.e.*, it is assumed that all microscopic pre-images of a given s contribute equally, and the back-mapped distribution is constructed as $\bar{p}(\sigma) = q(s)/2^{n-N}$. This corresponds to imposing $\ln \Omega = (n - N) \ln 2$ by construction and therefore isolates the contribution of finite sampling associated with the incomplete realisation of degenerate microscopic configurations. The resulting estimator is denoted by $S_{\text{map}}^{\text{inf}}$, where “inf” stands for the infinite-sampling limit.

The first benchmark system is defined by taking the reference single-site proba-

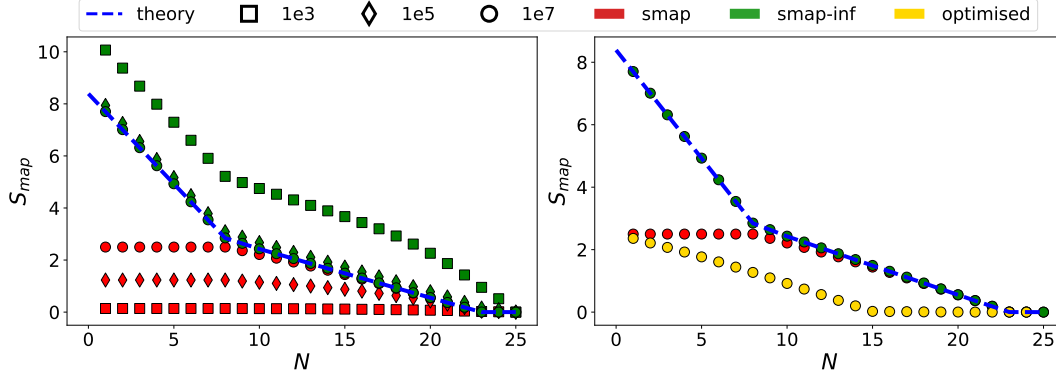


Figure 7.2: Mapping-entropy estimates for the pixel-based benchmark with near-deterministic sites. Minimum ME as a function of the number of retained spins N . The dashed blue line reports the theoretical prediction $S_{\text{map}}^*(N)$. **Left:** S_{map} (red) and $S_{\text{map}}^{\text{inf}}$ (green) evaluated on the fixed mapping prescribed by the theoretical ranking (cf. Figure 7.1), for three dataset sizes $P = 10^3, 10^5, 10^7$ (square/diamond/circle markers). **Right:** comparison at $P = 10^7$ between the same fixed-mapping estimates (red/green) and the value obtained by directly minimising the empirical ME over mappings via simulated annealing (yellow).

bilities $\{p_i\}$ from the pixel intensity profile of a image. The probability map, together with representative configurations sampled from the corresponding independent-spin model, is reported in Figure 7.1. As it can be observed, several pixels are associated with probabilities very close to unity, so that across the sampled configurations those degrees of freedom take the same value almost always.

This feature is captured quantitatively by the one-body entropy $s(p_i)$ defined in Equation 7.35. Pixels with $p_i \simeq 1$ contribute $s(p_i) \simeq 0$, as shown in Figure 7.1.c. Since, at fixed N , the optimisation of the ME reduces to selecting the $n - N$ discarded sites so as to maximise the sum of their one-body entropies (recall that this sum enters with an overall negative sign), the optimal strategy is to exclude from the sum the nearly frozen pixels and to discard instead the more uncertain ones, *i.e.*, those with p_i closer to $1/2$.

As a first benchmark, the theoretical prediction for the minimum ME $S_{\text{map}}^*(N)$ can be compared with the values obtained from the two standard, data-driven estimators introduced above, S_{map} and $S_{\text{map}}^{\text{inf}}$, while keeping the mapping *fixed* to the one selected by the theory (*i.e.*, the retention order shown in Figure 7.1). This comparison is reported in Figure 7.2.

Consider first the fully empirical estimator S_{map} . For small N a pronounced discrepancy emerges, essentially independently of the dataset size P , the empirical curve displays an extended plateau, while the theoretical prediction decreases already at the first retained sites. Only beyond $N \gtrsim 9$ does the agreement become satisfactory, and even then it is contingent on having a sufficiently large dataset; for smaller P a visible gap persists over the whole range of N . The origin of the initial mismatch can be traced back to the presence of nearly deterministic pixels (with $p_i \simeq 1$), which are precisely the first 9 sites selected by the theoretical ranking. In the theoretical construction, the degeneracy associated with discarded degrees of

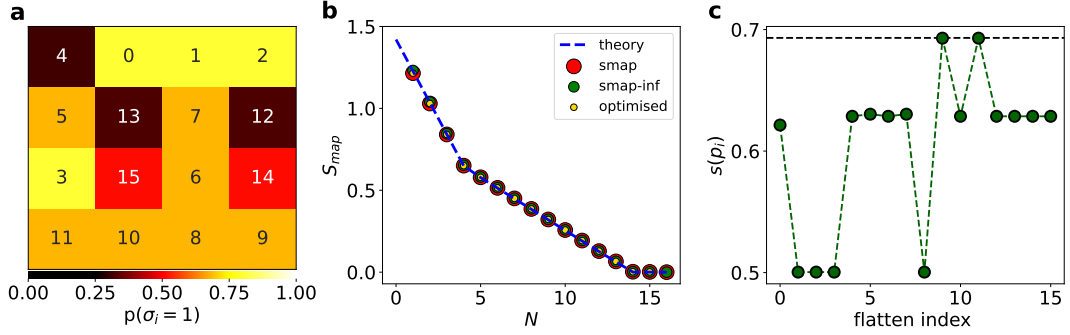


Figure 7.3: Rescaled $n = 16$ pixel benchmark (no near-deterministic sites), $P = 10^7$. (a) Reference single-site probabilities $p(\sigma_i = 1)$ after rescaling to the interval $[0.2, 0.8]$ (colour scale); the integer on each pixel indicates the order in which that pixel is selected as predicted by the theoretical ranking (0 is the first retained, then 1, etc.). (b) Minimum ME as a function of the number of retained spins N : theoretical prediction $S_{\text{map}}^*(N)$ (dashed blue) compared with the empirical estimators S_{map} (red) and $S_{\text{map}}^{\text{inf}}$ (green) evaluated on the fixed, theory-selected mapping, and with the value obtained by direct optimisation over mappings via simulated annealing (yellow). (c) One-body entropies $s(p_i)$ associated with the probabilities in panel (a), reported versus the flattened pixel index; the horizontal dashed line marks the maximum value $s(1/2) = \ln 2$.

freedom is taken to be 2^{n-N} , i.e., it is assumed that both microscopic values of each discarded spin are effectively realised. In finite samples, however, configurations that differ only in those nearly deterministic coordinates are typically absent, so that the empirical degeneracy is strongly reduced and the corresponding S_{map} does not follow the ideal scaling at small N .

This interpretation is corroborated by the comparison with $S_{\text{map}}^{\text{inf}}$. Since $S_{\text{map}}^{\text{inf}}$ enforces the ideal degeneracy term by construction, the discrepancy at small N is significantly reduced, and the agreement with $S_{\text{map}}^*(N)$ improves over the entire range, especially at large P . The residual deviations observed for the smallest datasets are consistent with finite-sampling noise in the empirical reconstruction of $p(\sigma)$ and $q(s)$, rather than with a breakdown of the theoretical ranking criterion.

The effect becomes even more transparent when the mapping is not fixed, but instead the ME is minimised directly over the space of mappings through simulated annealing (right panel of Figure 7.2). In this case the resulting values are systematically lower than those obtained by evaluating S_{map} on the theoretically selected mapping, and the associated pixel-selection order is qualitatively different. In particular, the set of pixels with $p_i \approx 1$ —which the theory would retain first because they have the smallest $s(p_i)$ —is retained only at much larger N by the data-driven optimisation. This inversion reflects the fact that, under finite sampling, discarding nearly deterministic variables can be advantageous: it reduces the impact of missing microscopic degeneracies and yields a smaller empirical S_{map} , even though it contradicts the ideal, infinite-sampling optimum.

The last benchmark considers a smaller version of the previous system, obtained by selecting a subset of pixels leading to a smaller configuration space ($n = 16$). In addition, the reference probabilities are rescaled such that $p_i \in [0.2, 0.8]$ for every

site (Figure 7.3.a). As a consequence, no degree of freedom is nearly deterministic: this is reflected by the corresponding one-body entropies $s(p_i)$, which remain bounded away from zero and one across the whole set of spins (Figure 7.3.c).

In this setting the comparison with the theoretical prediction changes qualitatively. For the largest data set ($P = 10^7$) all empirical curves collapse onto a single profile (Figure 7.3.b): the value obtained by evaluating S_{map} on the mapping prescribed by the theory, the estimator $S_{\text{map}}^{\text{inf}}$, and the fully optimised value obtained by simulated annealing are indistinguishable within numerical accuracy. This collapse indicates that the discrepancy observed in the previous benchmark is not a limitation of the analytical expression for $S_{\text{map}}^{\star}(N)$ per se, but rather a finite-sampling effect associated with an incomplete exploration of the microscopic configuration space. In particular, when near-deterministic degrees of freedom are present, the empirical degeneracy of CG states can be substantially smaller than the ideal value 2^{n-N} , which directly impacts the reconstruction of $\bar{p}(\sigma)$ and hence the ME.

Finally, it is worth noting that the theoretical framework can, in principle, be refined to account for this effect by relaxing the assumption of a uniform degeneracy $\Omega(s) = 2^{n-N}$ and allowing instead for an effective, state-dependent degeneracy when some variables are sampled only in one of their microscopic states. This extension makes it possible to capture the finite-sampling reduction of the degeneracy term, and therefore to describe qualitative effects that are absent in the ideal theory, such as the appearance of extended nearly flat regions in the minimum-ME curves and possible shifts of the optimum toward higher resolution. A first discussion in this direction is reported in Appendix E.

7.4 Conclusions

This chapter has developed a systematic route to approximate and interpret the Mapping Entropy by means of a multi-body decomposition. Starting from the exact expression of the ME in terms of the microscopic distribution $p(\sigma)$ and the back-mapped distribution $\bar{p}(\sigma)$, it has been rewritten in a form that naturally separates contributions of increasing order in the correlations of the retained degrees of freedom. This construction provides both a conceptual interpretation of the ME and a practical strategy for its approximation. In particular, the ME can be understood as the sum of a degeneracy contribution and a hierarchy of information-theoretic corrections. Moreover, the expansion can be truncated at a chosen order to obtain controlled approximations of the full ME.

In the special case of independent spins, the formalism becomes fully explicit. The factorised structure of $p(\sigma)$ implies a closed-form expression for the ME, and the optimisation at fixed number of retained sites reduces to a simple ranking problem: the optimal mapping is obtained by retaining the spins with the smallest one-body entropy $s(p_i)$. This yields an analytical prediction for the minimum ME $S_{\text{map}}^{\star}(N)$ and clarifies its monotonic dependence on the number of retained degrees of freedom.

In addition, the role of the resolution (and of the associated degeneracy factor) has been made explicit by linking S_{map} to the size of the microscopic pre-image of each **CG** configuration, thereby providing a clear interpretation of how a change in resolution affects the information loss.

The analytical predictions have been validated numerically by analysing synthetic datasets generated from an independent-spin model defined by pixel-based probabilities. Importantly, the empirical evaluation has been performed under the same information constraints encountered in practical applications: probabilities have been reconstructed exclusively from finite samples, without exploiting the known analytical form of the generating distribution, and two standard estimators have been compared, S_{map} and $S_{\text{map}}^{\text{inf}}$, differing only in the treatment of the degeneracy factor. In the first benchmark, the presence of probabilities close to 1 produces an extended discrepancy between the theoretical curve and the fully empirical estimator at small N , together with a qualitative change in the mapping selected by direct optimisation. The smaller $n = 16$ benchmark system, obtained by rescaling the probabilities to $p_i \in [0.2, 0.8]$, removes this effect and leads to an essentially perfect collapse of all estimates for sufficiently large sample sizes. These results demonstrate that the theoretical framework accurately captures the independent-spin optimum, and that the dominant deviations observed in the first benchmark originate from finite-sampling effects that hinder the accurate estimate of the **CG** states degeneracy when nearly deterministic degrees of freedom are present.

The natural next step is to move beyond the independent-spin setting and consider interacting systems, where the correlation terms of the expansion become non-trivial. In that case, the practical question is whether low-order truncations of the multi-body decomposition provide accurate approximations of the full **ME** and, crucially, whether they predict mappings close to those obtained by minimising the exact **ME**. Establishing the accuracy of the first non-trivial orders and identifying regimes where higher-order terms become necessary will determine the operational value of the decomposition as a tractable optimisation tool for realistic, interacting models.

Take-home message

Mapping Entropy quantifies the information loss associated with replacing a high-resolution description by a reduced one. In this chapter, this information loss was decomposed into contributions of increasing statistical order, separating the effect of mapping degeneracy from the effect of correlations among microscopic variables. The resulting framework clarified which parts of the atomistic probability structure are lost under **Coarse-Graining** and provided a controlled reference case in which, for non-interacting systems, the **Mapping Entropy** reduces to a sum of independent one-body contributions.

8

Neural Network Configuration Space

8.1 Introduction

The study of neural networks has long been intertwined with statistical physics, with concepts originally developed for disordered systems providing a powerful framework to understand learning processes. Since the seminal work of Gardner and Derrida, training a neural network has been formulated in terms of an energy landscape defined over the space of parameters, where learning corresponds to identifying weight configurations that minimize a loss function, in close analogy with the search for ground states in physical systems with quenched disorder (Gardner et al., 1988; Gardner, 1988).

This analogy has led to major theoretical advances. Classical approaches based on the statistical mechanics of disordered systems and spin-glass theory have clarified the ability of simple neural architectures to memorize and generalize (Seung et al., 1992; Watkin et al., 1993; Mézard et al., 1987; A. Engel et al., 2001). Extensions of this framework have also related geometric properties of the loss landscape, such as flatness or the presence of dense regions of solutions, to the generalization ability of trained models (Baldassi et al., 2015; Baldassi et al., 2016; Baldassi et al., 2016; Baldassi et al., 2021). However, many of these results rely on simplifying assumptions, such as infinite system size or unstructured input data, which are rarely met in realistic learning problems.

Recent work has started to overcome these limitations by explicitly incorporating data structure into the theory. In particular, second-order statistics of the in-

This chapter is based on material originally published in M. Mele, R. Menichetti, A. Ingrosso, and R. Potestio, *Density of states in neural networks: an in-depth exploration of learning in parameter space*, Transactions on Machine Learning Research (2025). The material has been revised and adapted for consistency with the structure and notation of this thesis, and is reproduced in accordance with the terms of the applicable licence. The published version is available at: <https://openreview.net/forum?id=BLDtWlFKhn>

puts have been used to move beyond featureless random patterns (Opper et al., 2001; Shinzato et al., 2008; Ingrosso, 2020), while analytically tractable models based on Gaussian data and mixtures of Gaussians have provided a controlled setting to study the effect of structured datasets on learning and generalization (Mignacco et al., 2020; Gerace et al., 2021; Gerace et al., 2024; Loureiro et al., 2021b; Loureiro et al., 2021a; Loureiro et al., 2023; Cornacchia et al., 2023). More recently, higher-order statistics have also been shown to play a central role in the emergence of internal representations and learned features (Ingrosso et al., 2022; Belrose et al., 2024). These developments mark an important shift toward a more realistic theoretical description of neural networks trained on structured data.

Despite this progress, most studies still focus on a limited portion of parameter space, namely the subset of configurations that minimize, or nearly minimize, the loss. While this viewpoint is natural from an optimization perspective, it leaves aside the broader organization of the loss landscape. Yet, the way all possible configurations are distributed across loss values can itself provide valuable information on the learning problem.

In this chapter, we adopt this complementary perspective by investigating the *Density of States* (DoS) of a neural network, namely the number of parameter configurations associated with a given loss value. This quantity, central in statistical physics, provides a global characterization of the parameter space and of its dependence on the training data.

To estimate it, we employ the *Wang-Landau* (WL) algorithm, an enhanced-sampling method originally introduced to reconstruct the DoS of complex physical systems (F. Wang et al., 2001b; F. Wang et al., 2001a; Yin et al., 2012; Vogel et al., 2014; Farris et al., 2021). By adaptively estimating the microcanonical entropy of the loss, this method enables an efficient exploration of parameter space, including regions that would be difficult to access with conventional sampling strategies.

Within this framework, we analyse neural networks with discrete weights and study how the statistical and geometrical properties of the training set shape the resulting DoS. We show that structured datasets leave distinct signatures on the loss spectrum, and that several of these features can be reproduced and rationalized through controlled synthetic data. In this way, the DoS provides a novel perspective on the interplay between data structure and network behaviour, complementing more standard analyses based on optimization and generalization.

8.2 Materials and Methods

8.2.1 Neural Network Architectures

In this chapter, we consider finite-size feed-forward neural networks with binary synaptic weights, and study their configuration space through the loss function defined over all possible parameter assignments. Since the general formalism of feed-

forward networks and supervised learning has already been discussed in [Chapter 6](#), here we introduce only the specific architectures analysed in the present work.

We begin from the simplest case, namely a single-layer binary neural network, or perceptron, which maps an input vector $\xi \in \mathbb{R}^N$ to a binary output according to

$$o(\xi; \mathbf{W}) = \text{sgn}(\mathbf{W} \cdot \xi), \quad (8.1)$$

where $\mathbf{W} \in \{-1, 1\}^N$ is the synaptic weight vector. Given a set of $P = \alpha N$ patterns $\{\xi^\mu\}_{\mu=1}^P$ and their corresponding labels $\{\sigma^\mu\}_{\mu=1}^P$, with $\sigma^\mu \in \{-1, 1\}$, the learning problem consists in identifying a weight configuration $\hat{\mathbf{W}}$ such that $o(\xi^\mu; \hat{\mathbf{W}}) = \sigma^\mu$ for all μ . In statistical-mechanical terms, this problem can be expressed through the energy function

$$E(\mathbf{W}) = \sum_{\mu=1}^P \Theta(-\sigma^\mu \mathbf{W} \cdot \xi^\mu), \quad (8.2)$$

where $\Theta(x)$ is the Heaviside step function. This quantity counts the number of incorrect input-output associations produced by a perceptron with weight vector $\mathbf{W}$, and therefore plays the role of the loss function in the present discrete setting ([Gardner et al., 1988](#); [Gardner, 1988](#); [Seung et al., 1992](#); [Watkin et al., 1993](#)).

Because of the binary nature of the weights and the use of the sign activation function, the perceptron energy satisfies the symmetry

$$E(\mathbf{W}) = P - E(-\mathbf{W}), \quad (8.3)$$

which implies that the number of states at energy levels equidistant from the middle of the spectrum is identical by construction. As discussed in the following sections, this property is directly reflected in the shape of the density of states for binary classification tasks.

To move beyond linearly separable decision boundaries, we also consider a one-hidden-layer network with binary weights. In this case, an input vector $\xi \in \mathbb{R}^N$ is mapped to a hidden layer of N_h neurons through a weight matrix $\mathbf{W}^{(1)} \in \{-1, 1\}^{N_h \times N}$, yielding hidden activations

$$h_k = \text{sgn}\left(\sum_{i=1}^N W_{ki}^{(1)} \xi_i\right), \quad k = 1, \dots, N_h. \quad (8.4)$$

The hidden representation is then mapped to the output layer through a second weight matrix $\mathbf{W}^{(2)} \in \{-1, 1\}^{N_o \times N_h}$, so that the final output reads

$$\mathbf{o} = \phi(\mathbf{W}^{(2)} \mathbf{h}), \quad (8.5)$$

where ϕ denotes either the sign function, in the binary classification case, or the argmax rule for multi-class classification.

For a dataset of P labelled patterns, the corresponding training energy is defined as

$$E(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}) = \sum_{\mu=1}^P \left[1 - \delta \left(\sigma^{\mu}, \phi \left(\mathbf{W}^{(2)} \operatorname{sgn}(\mathbf{W}^{(1)} \boldsymbol{\xi}^{\mu}) \right) \right) \right], \quad (8.6)$$

with δ the Kronecker delta. Also in this case, learning amounts to identifying weight configurations that minimize the number of classification errors. Compared with the perceptron, the hidden layer introduces an internal representation of the input and allows the network to implement non-linear decision rules in the original input space.

Symmetries also exist in the one-hidden-layer network, although binary and multi-class classification must be distinguished. In the binary case, the system retains a symmetry with respect to the middle of the energy spectrum. Indeed, if both weight matrices are simultaneously negated, the energy remains unchanged,

$$E(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}) = E(-\mathbf{W}^{(1)}, -\mathbf{W}^{(2)}). \quad (8.7)$$

By contrast, negating only one of the two matrices maps a configuration to another one whose energy is symmetric with respect to the centre of the spectrum, namely

$$E(-\mathbf{W}^{(1)}, \mathbf{W}^{(2)}) = E(\mathbf{W}^{(1)}, -\mathbf{W}^{(2)}) = P - E(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}). \quad (8.8)$$

This property follows directly from the binary nature of the sign output activation function. When the output is instead defined through an argmax operation, as in the multi-class case, this symmetry is generally lost.

8.2.2 Datasets

To investigate how the statistical properties of the training data affect the DoS of neural networks, we consider a set of learning tasks based on both real-world and synthetic datasets.

The first scenario involves real-world datasets. The classification task is performed for various sizes of the training set $\alpha = P/N$, where α is gradually increased from 0.1 to 1.0 by randomly adding examples from the two digit classes considered. As a result, for a given total number of inputs P , there might be a discrepancy in the number of examples between the two classes ($P_0 \neq P_1$): this condition is commonly referred to as an “unbalanced” dataset. By contrast, a “balanced” situation entails equally populated classes ($P_0 = P_1$). In the main text, we show results obtained on the classic MNIST dataset (LeCun et al., 1998), focusing specifically on binary classification problems involving only the first two classes (0 and 1). Additional classification problems on the MNIST dataset are reported in Appendix F, along with results obtained using Fashion-MNIST (Xiao et al., 2017).

In the second scenario, we use a series of approximations of an original real-world dataset that retain selected statistical properties of each class. In particular,

we construct Gaussian clones of the datasets, by sampling inputs from a mixture of Gaussians each capturing the mean vector and covariance matrix of the images in that class. We call this configuration GM (Gaussian mixture), where inputs possess the correct mean and covariance per class but where all higher-order cumulants are zero. We also compare the results on a simplified isotropic Gaussian mixture (2isoGM), where only the mean of each class has been preserved, and the pixel-wise variance of the two classes is equal to the geometric mean of the real variances.

Finally, the last two scenarios involve *synthetic* datasets. We collect hereafter a few basic information on the data-generating process; all technical details can be found in [Appendix F](#). In the first case, each class represents an isotropic distribution of points in N dimensions. Two controlling parameters are employed: the inter-class separation, defined as the norm of the difference between the mean vectors of the two classes, and the angle between these two mean vectors. In the second case, we consider a perceptron learning problem with randomly generated data in a [Teacher-Student \(TS\)](#) setup. Each input component ξ_i^μ is drawn i.i.d. at random with probability $1/2$ in $\{-1, 1\}$, and the outputs σ^μ are generated using a *teacher* vector W_T as $\sigma^\mu = \text{sgn}(W_T \cdot \xi^\mu)$, implying that data points are linearly separable by construction as the solution W_T always exists.

8.2.3 Wang-Landau Sampling of the Parameter Space

The [WL](#) algorithm and its general theoretical framework have been discussed extensively in [Chapter 6](#). Here, we therefore limit the discussion to the implementation details specific to the present application, the estimation of the loss [DoS](#) in finite neural networks with binary weights.

The quantity of interest is the [DoS](#)

$$\Omega(E) = \sum_{\mathcal{W}} \delta(E(\mathcal{W}) - E), \quad (8.9)$$

namely the number of network configurations associated with a given value of the training loss. Its direct evaluation is computationally prohibitive because the size of the configuration space grows exponentially with the network size. A perceptron with N binary weights has 2^N possible configurations, while a one-hidden-layer network with N input neurons, N_h hidden neurons, and N_o output neurons has $2^{N_h N + N_o N_h}$ configurations ([F. Wang et al., 2001a](#); [F. Wang et al., 2001b](#)). As a concrete example, a perceptron with $N = 100$ binary weights has $2^{100} \simeq 1.3 \times 10^{30}$ possible configurations; even assuming that a previously unvisited state could be sampled on average every nanosecond, exploring the full configuration space would require more than 10^{13} years. In this regime, [WL](#) sampling provides an efficient strategy to reconstruct $\Omega(E)$ through a sequence of biased random walks in parameter space ([F. Wang et al., 2001b](#); [F. Wang et al., 2001a](#); [Shell et al., 2002](#)).

In practice, the simulation is performed in the entropy representation, $S(E) =$

$\ln \Omega(E)$, and follows the standard histogram-flatness version of the **WL** algorithm. At each iteration, the current estimate of the entropy is updated together with the visitation histogram in energy space. A Monte Carlo move consists in proposing a new network configuration by flipping a subset of the binary weights; if $\mathcal{W}$ and $\mathcal{W}'$ denote the current and proposed configurations, with corresponding energies $E = E(\mathcal{W})$ and $E' = E(\mathcal{W}')$, the move is accepted according to the usual **WL** rule

$$A(\mathcal{W} \rightarrow \mathcal{W}') = \min \left\{ 1, \frac{\Omega(E)}{\Omega(E')} \right\} = \min \left\{ 1, \exp[-(S(E') - S(E))] \right\}. \quad (8.10)$$

Once the histogram satisfies the flatness criterion, namely when the population of each visited energy bin is at least 80% of the average histogram count, it is reset and the refinement parameter is reduced according to the standard schedule $F_{k+1} = F_k/2$. The simulation is stopped when $F < 10^{-6}$. The resulting **DoS** is determined up to an overall multiplicative constant, which can be fixed in the present case because the full energy spectrum is sampled and the total number of admissible configurations is known exactly. The normalization is therefore obtained by imposing that the sum of the estimated **DoS** over all energies equals the total number of network configurations. Additional implementation details of the **WL** algorithm, including validation tests, convergence analysis, and scaling behaviour, are reported in [Appendix F](#).

8.3 Results

In most studies of neural networks, the focus lies on parameter configurations that minimise the loss function. Here, by contrast, the analysis is extended to the full loss spectrum, investigating how the number of parameter configurations associated with each value of the loss – that is, the **DoS** – depends on the properties of the network and its training set. Several interconnected factors, among which network architecture, activation functions, and statistical structure of the input, play a critical role in shaping the loss **DoS**. Rationalizing this interplay bears twofold consequences: on the one hand, it provides insight into the properties of the dataset itself; on the other hand, it can help us in understanding the inner working of task-optimized neural networks. In the following, we report the results of our work in one of these directions, our main goal being to *elucidate how data structure impacts the **DoS***.

In the first part of this section, we apply our method to a benchmark dataset commonly employed in supervised learning. We examine both a perceptron and a one-hidden layer architecture solving classification tasks. A drawback in the theoretical analysis of real-world datasets is that the probability distribution underlying the input data is generally not known: to make sense of our results, we thus turn to analyse synthetic datasets whose generative model is controllable with a small set

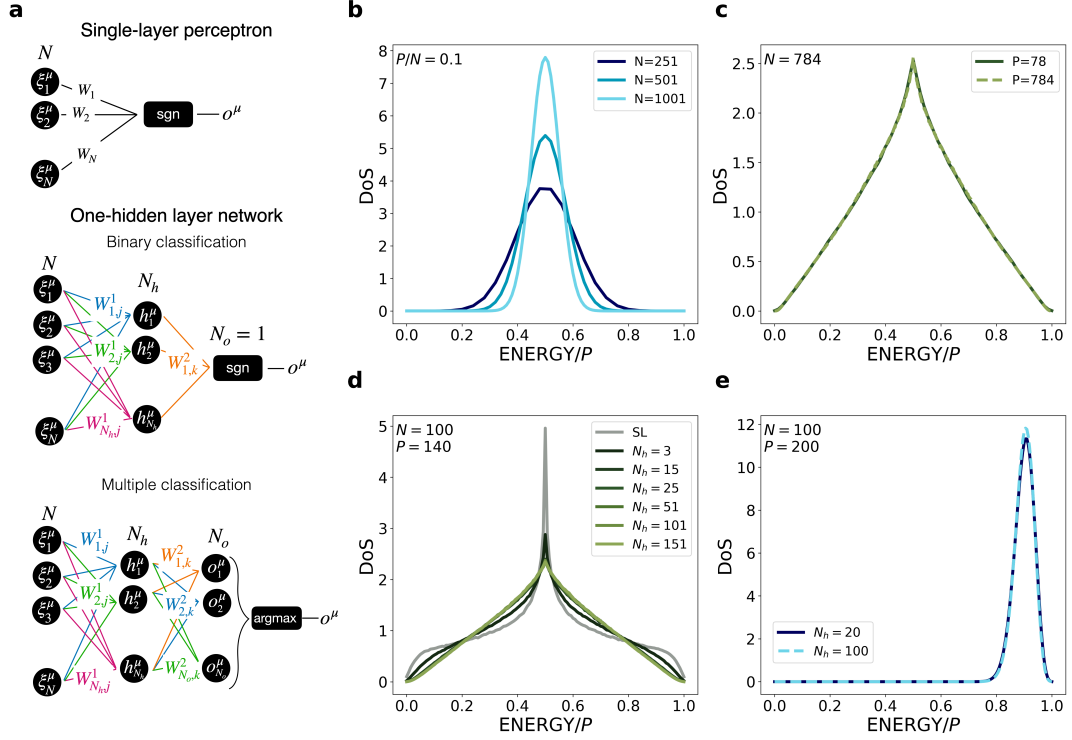


Figure 8.1: Comparative analysis of the DoS in classification tasks of random or structured data across different network architectures and input dimensions. (a) Schematic representations of the network architectures: single-layer perceptron (top), and one-hidden layer network for both binary classification (middle) and multiple classifications (bottom), illustrating the input layer (N), hidden layer (N_h), and output layer (N_o), along with their respective weight connections. (b) DoS for binary classification of random input data, in teacher-student setup, using a single-layer perceptron. (c) DoS for binary classification of the first two classes of the MNIST dataset with $N = 784$ input dimensions by single-layer perceptron, illustrating results for different learning complexities ($P/N = 1.0$ and $P/N = 0.1$). (d) DoS for binary classification of the first two classes of the MNIST dataset with $N = 100$ input dimensions, with a fixed data size of $P = 140$ and varying network architectures, including a single-layer (SL) perceptron and one-hidden layer network with different numbers of hidden layer neurons (N_h). (e) DoS for multi-class classification of the MNIST dataset with $N = 100$ input dimensions, showing results for varying numbers of hidden neurons (N_h) in the one-hidden layer network.

of parameters, recovering the same phenomenology observed in the case of realistic input data.

8.3.1 Density of States Analysis of Benchmark Tasks

We start off addressing the problem of computing the DoS for simple networks applied to both random and structured input data. We analyse both a binary, single-layer (SL) perceptron and a network with one hidden layer with varying size. The latter architecture is employed for binary classification as well as multi-class labelling.

Figure 8.1.a shows the DoS for the aforementioned architectures and classification tasks, demonstrating how various factors modify its properties. The DoS in perceptron solving a binary classification task on both random (Figure 8.1.b)

and real data (Figure 8.1.c) is displayed. We also show the results obtained in a one-hidden layer neural network with an increasing number of neurons N_h in the hidden layer, for binary (Figure 8.1.d) and multi-class classification on the MNIST dataset (Figure 8.1.e)

Figure 8.1.b,c,d illustrate that, in the case of binary classification, the DoS curves are symmetric with respect to the mid-energy value ($E = 0.5$) irrespective of the network architecture. As anticipated, the symmetry is induced by the sign activation function in conjunction with the binary nature of the weights, *i.e.*, $W = \{-1, 1\}^N$. Such symmetry is lost in the presence of multi-class classification with an argmax function (see Figure 8.1.e), where instead the DoS peaks at a value of the energy corresponding to the typical error $1 - 1/C$ obtained when guessing C possible classes with uniform probability.

We now compare the characteristics of the DoS curves for binary classification *via* a perceptron in two distinct scenarios: i.i.d. data labelled by a teacher network (Figure 8.1.b) and structured data, exemplified by the classification of the first two MNIST classes (Figure 8.1.c). Both curves exhibit a maximum at the central energy value ($E = 0.5$), indicating that most network configurations incorrectly label half of the input data. However, the shape of the distribution is markedly affected by the type of data used in the binary classification. For i.i.d. data, the DoS follows a Gaussian distribution centred at 0.5 (Appendix F). For structured data, the DoS decreases almost linearly away from the centre of the energy range.

The non-Gaussian shape of the DoS is maintained even when the network architecture is modified by adding a hidden layer (Figure 8.1.d). Specifically, we used a one-hidden layer network with varying inner layer sizes to perform binary classification of the first two MNIST classes (reshaped into $N = 100$ dimensions, *i.e.*, 10×10 instead of the original 28×28). We note that the general properties of the curve, symmetries and shape, are preserved. The DoS obtained with a perceptron, taken as a reference, is different from the one obtained with a dataset of larger-sized images; in particular, this curve has a sharper central peak, from which a slower decrease is observed while moving towards lower and higher energy values. This notwithstanding, as the number of hidden layers is increased, the DoS curve gradually converges to a shape qualitatively much more similar to the one observed in the case of Figure 8.1.c (SL perceptron, MNIST dataset with images of size 28×28).

Taken together, these findings suggest that the shape of the DoS is influenced by the architecture of the network as well as by the structure of the input data, in a non-trivial interplay between the two. In order to better understand the origin of specific properties of the DoS, we thus modulated some features of the underlying data and investigated the relationship these bear with the shape of the resulting curve.

Firstly, we studied the effect of class imbalance on MNIST dataset, *i.e.*, the discrepancy between the number of data points $\{P_0, P_1\}$ in the two classes, while keeping the total number of points $P = P_1 + P_0$ fixed. To assess the impact of class im-

balance on the dataset, the DoS curves of the binary classification were computed for different values of P_1 , thus including scenarios where class 1 was either predominant ($P_1 > P_0$) or in the minority ($P_1 < P_0$). Figure 8.2.b shows that as class imbalance is increased, the peak of the DoS shifts away from the centre of the spectrum.

Irrespective of which class is predominant, we observe the appearance of two peaks that shift from the centre as the imbalance increases, due to the symmetry of the DoS curve. When P_1 is significantly greater than P_0 (green curves), the DoS shows pronounced peaks away from the centre, indicating a higher concentration of states at low/high energy values. As P_1 approaches $P/2$ (red curve) from above, the DoS curves become more centralized, indicating a more balanced distribution of states across the energy spectrum. When P_1 is less than $P/2$ (and thus less than P_0 , blue curves), a pattern similar, albeit mirrored, to the first case emerges, with the peaks moving away from the centre as P_1 is decreased.

This behaviour can be understood by noting that the two symmetrical peaks in the DoS correspond to distinct sets of solutions arising from the class imbalance. The low-energy peak represents configurations where the decision boundary optimally classifies the majority class, while the minority class contributes most of the errors (and vice versa for the high-energy peak). This duality is a natural consequence of the binary weights of the perceptron and the definition of energy, which symmetrically measures the misclassifications of the two classes.

The shifting of the peaks described above is clearly reflected in a scatter plot (Figure 8.2.c) that correlates the position of the low-energy maximum of the DoS curve with the degree of unbalancing, computed as by $|0.5 - P_1/P|$. The plot shows results for two distinct values of dataset size, $\alpha = 0.1$ (squares) and $\alpha = 1$ (diamonds). As the degree of imbalance increases, the peak energy shifts proportionally, thereby providing a quantifiable measure of the impact of class distribution on the DoS.

These results show how class imbalance plays a crucial role in shaping the DoS, in a way that is further dependent on the structure of the input data; most importantly, class imbalance does not change the location of the DoS peak for randomly distributed input: we observe such behaviour only in the presence of structured data. This conclusion is supported by further analyses we carried out on the MNIST and Fashion-MNIST datasets (see Appendix F).

8.3.2 Density of States Analysis of Synthetic Tasks

To further elucidate the factors impacting the DoS, we studied simple classification problems on *synthetic* datasets, designed in such a way that a systematic study can be carried out over a small number of control parameters.

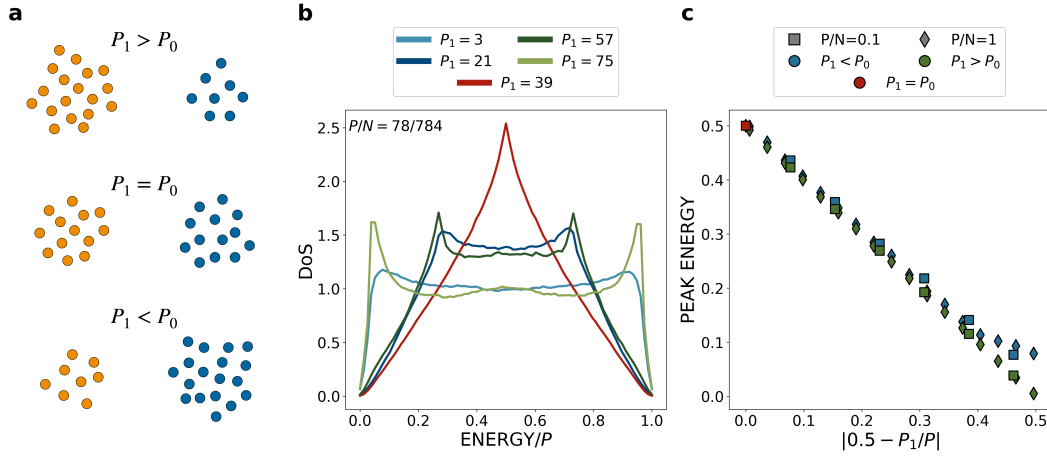


Figure 8.2: *DoS* analysis for binary classification of MNIST digits (0 and 1) under various class imbalances. (a) Schematic representation of class imbalance. In the top section of the panel, $P_1 > P_0$, there are more elements in class 1 (orange) than in class 0 (blue). In the middle section, $P_1 = P_0$, the classes are balanced. In the bottom section, $P_1 < P_0$, there are fewer elements in class 1 than in class 0. The total number of points $P_1 + P_0 = P$ is fixed. (b) *DoS* at a fixed learning complexity of $P/N = 78/784$ resulting in MNIST binary classification for different class imbalances. The legend indicates the number of elements in class 1 (P_1). Perfect class balance is achieved when $P_1 = P_0 = P/2 = 39$ (red curve). Larger deviations from this value indicate greater class imbalance. Blue curves represent a predominance of class 0, while green curves represent a predominance of class 1; the lighter the colour the greater the class imbalance. (c) Peak energy values plotted against the absolute difference $|0.5 - P_1/P|$, showing the correlation between peak energy and class imbalance. This correlation does not depend on which class is predominant: blue points indicate class 0 is predominant, and green points indicate class 1 is predominant. Results are shown for two values of learning complexity: $P/N = 0.1$ (squares) and $P/N = 1.0$ (diamonds).

Density of States in Classification Problems with Data From Mixtures of Isotropic Gaussians

Here, we focus on the *DoS* obtained from input data sampled from Gaussian probability densities and study its changes as we modify specific parameters of these distributions. We construct a binary classification task where the different classes consist of isotropic clouds of points in a high-dimensional space (we set $N = 41$). The distributions for each class are Gaussians with zero mean and unit variance (further details can be found in supplementary material Section I). We systematically vary the inter-class distance $\Delta\mu$, defined as the distance between the mean points of the distributions, as well as the angle θ between the mean vectors.

Figure 8.3.b shows the *DoS* curves obtained by varying the value of $\Delta\mu$, in the simple case of $\theta = 180^\circ$. We expect that a large value of $\Delta\mu$ is associated to rather trivial classification tasks, since the two groups of points are nearly distinct and, consequently, a great number of planes in input space can successfully separate them. In contrast, a $\Delta\mu$ close to zero corresponds to overlapping clouds: in this latter case, no plane likely exists that allows a distinction between them.

This expectation is indeed confirmed by the behaviour of the *DoS*: the latter is bell-shaped, with a peak at the centre of the spectrum when $\Delta\mu$ is small, implying that the vast majority of network parameter configurations wrongly classify the

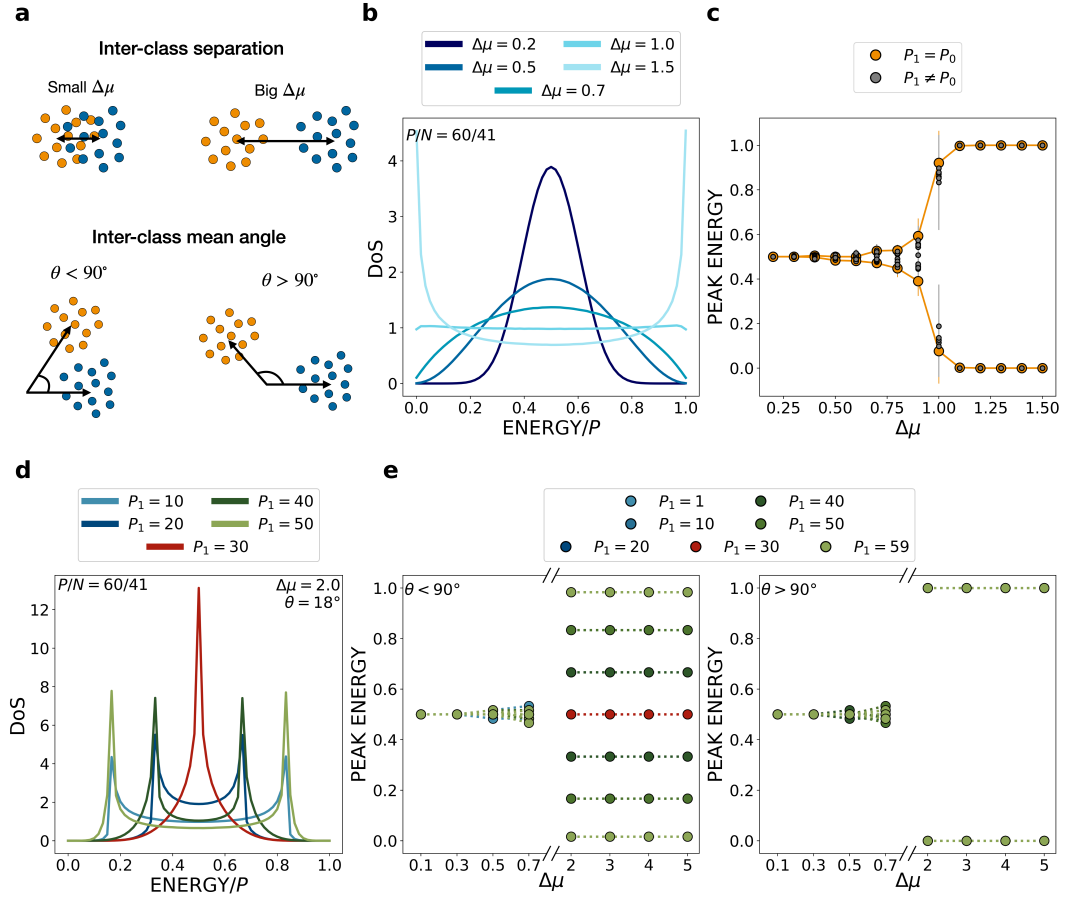


Figure 8.3: DoS analysis for binary classification of synthetic datasets studying the impact of inter-class separation ($\Delta\mu$) and inter-class mean angle (θ). The angle between the two mean vectors uniformly ranges from 180° to 18° . (a) Pictorial representations of inter-class separation ($\Delta\mu$) at the top and inter-class angle (θ) at the bottom. (b) DoS at a fixed learning complexity of $P/N = 60/41$ obtained in the binary classification for multiple inter-class separations $\Delta\mu$ and fixed inter-class angle ($\theta = 180^\circ$) for balanced classes. The legend indicates that lighter colours correspond to larger inter-class distances. (c) Peak energy values plotted against the inter-class distance $\Delta\mu$ at $\theta = 180^\circ$ for balanced classes (orange) and imbalanced classes (gray). The plotted values represent the average from 20 independent replicas of the system in the same configuration, with associated statistical deviations. (d) DoS at a fixed learning complexity of $P/N = 60/41$ obtained in the binary classification for large inter-class separations ($\Delta\mu = 2.0$) and small inter-class mean angle ($\theta = 18^\circ$) for different class imbalances. The legend indicates the number of elements in class 1 (P_1). Perfect class balance is achieved when $P_1 = P_0 = P/2 = 30$ (red curve). Larger deviations from this value indicate greater class imbalance. Blue curves represent a predominance of class 0, while green curves represent a predominance of class 1; lighter colours indicate greater class imbalance. (e) Peak energy values plotted against the inter-class distance ($\Delta\mu$) for different values of class imbalance as reported in the legend. Red points denote balanced classes. Blue and green points indicate unbalanced classes with a predominance of class 0 and 1, respectively; lighter colours indicate greater imbalance. The two panels differ by the value of the inter-class mean angle, with $\theta < 90^\circ$ on the left and $\theta > 90^\circ$ on the right.

inputs 50% of the times. As $\Delta\mu$ increases, the DoS broadens significantly: for sufficiently high values of $\Delta\mu$ the DoS develops two distinct peaks at the lowest and highest energy values, implying that most parameter sets correspond to classification planes that correctly label the data (or, alternatively, that make the wrong labelling all the times, which is equivalent to perfect labelling up to a sign).

In Figure 8.3.c we show the position of the DoS peaks as a function of $\Delta\mu$ ($\theta = 180^\circ$). The orange curve represents the balanced case ($P_1 = P_0$), while the grey points correspond to scenarios with class imbalance ($P_1 \neq P_0$). In this context, class imbalance does not affect the location of the DoS peaks: the trend is indeed consistent for both balanced and unbalanced scenarios.

We then extend our analysis to a synthetic dataset where both the angle and inter-class distance are varied systematically; we focus in particular on the two different regimes of small and large inter-class distances. Results are summarized in Figure 8.3.d,e. For large inter-class distances (e.g. $\Delta\mu = 2$) and low angles (e.g. $\theta = 18^\circ$), the DoS curves exhibit several properties observed in real datasets (Figure 8.2.b): the curves lose their typical bell shape, with the location of the maximum density being controlled by class imbalance. When $P_1 = P/2 = 30$, the DoS (red curve) displays a single peak at the centre of the energy spectrum. As P_1 deviates from $P/2$, with larger values marked in green and smaller values in blue, the peak shifts away from the centre. The peak location approaches the tail of the spectrum as P_1 deviates further from the balanced case.

The location of the maximum DoS reveals two distinct regimes based on the angle between the mean vectors: $\theta > 90^\circ$ and $\theta < 90^\circ$. In the first regime, class imbalance does not affect the peak location of the DoS curve. In contrast, for smaller angles, it plays a crucial role in determining the position of the peak, which shifts systematically towards the edges as the imbalance increases. The influence of class imbalance on the peak location is diminished when the inter-class separation is small. For small $\Delta\mu$, the maximum of the DoS remains centred in the energy spectrum, regardless of the angle between the mean vectors.

The results discussed thus far demonstrate that synthetic datasets have considerable power in replicating the behaviour observed in real-world cases. This provides valuable insights in the way neural networks carry out classification tasks: particularly, in the context of binary labelling involving MNIST data, we observe that the angle between the inter-class mean vectors generally falls below 90° (see Appendix F); such observation aligns with the results obtained on synthetic data, where class imbalance significantly influences the location of the peak of the DoS curves only for small angles, as it is the case for MNIST data.

Density of States in Classification Problems with Mislabelled Data

Label noise can affect the training performance and robustness of neural networks. Here we investigate how the DoS is affected by the presence of *mislabelling*, i.e., the

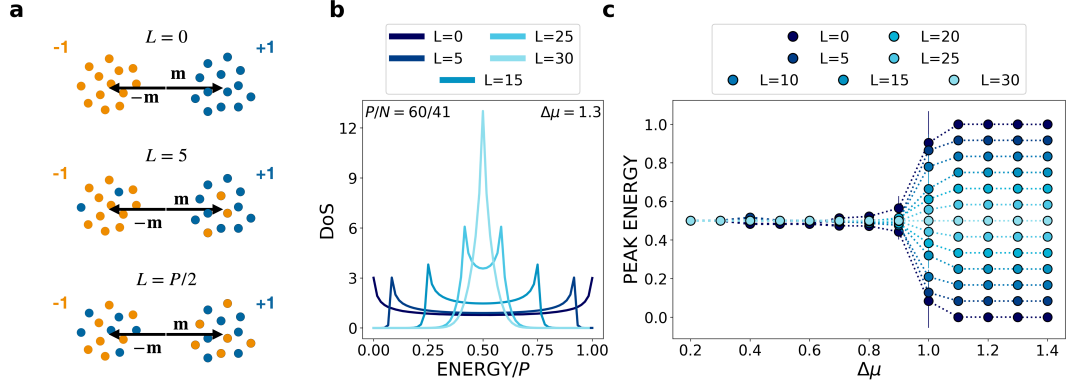


Figure 8.4: *DoS analysis for binary classification of synthetic datasets examining the impact of mislabelled elements at different inter-class separations $\Delta\mu$. (a) Pictorial representation of feature mislabelling for different numbers of mislabelled elements L . Each class is represented by an isotropic cloud of points, with one class having a mean vector $\mathbf{m}$ labelled as +1 (blue) and the other class with mean vector $-\mathbf{m}$ labelled as -1 (orange). The total number of points remains constant across different values of L . In the absence of feature noise (i.e., $L = 0$), there are an equal number of points in each class. Mislabelling is introduced by randomly selecting L elements and flipping their labels. For $L = 0$, all points are correctly labelled, while for increasing L (e.g., $L = 5$ and $L = P/2$), a portion of points from the cloud labelled +1 are relabelled as -1, and a portion of points from the cloud labelled -1 are relabelled as +1, leading to label noise. (b) DoS at a fixed learning complexity $P/N = 60/41$ and inter-class separation ($\Delta\mu = 1.3$), obtained in the binary classification for various numbers of mislabelled elements (L). The legend indicates that lighter colours correspond to a higher number of mislabelled elements (greater L). (c) Peak energy values plotted against the inter-class distance for different numbers of misclassified elements. The plotted values in the range $\Delta\mu \in [0.8, 1.3]$ represent the average from 10 independent replicas of the system in the same configuration, with associated statistical deviations.*

attribution of a given data point to the wrong class, which we can easily model in a synthetic dataset by switching the label of subsets of points generated from a mixture of Gaussians.

Specifically, we analyse a setup where some input patterns in the dataset are mislabelled. Each class is an isotropic Gaussian cloud of points, with the two mean vectors identifying the classes as antiparallel ($\theta = 180^\circ$); one cloud has a mean vector $\mathbf{m}$ and is labelled as +1, while the other cloud has a mean vector $-\mathbf{m}$ and is labelled as -1. We vary both the inter-class separation $\Delta\mu$ and the number of mislabelled elements L . To achieve the latter, a set of elements L is chosen at random, and their labels are reversed so that points from the cloud with mean $\mathbf{m}$ are labelled as -1, and/or points from the cloud with mean $-\mathbf{m}$ are labelled as +1. The parameter L ranges from 0, indicating all points are correctly labelled, to $P/2$, representing the maximum degree of disorder where half of the points are mislabelled.

Figure 8.4 shows that, as in previous scenarios, the peak of the DoS remains centrally located for any L value when the classes are close to each other ($\Delta\mu < 1$). However, as the inter-class separation increases, the introduction of mislabelled elements causes the peaks to shift. Notably, in the absence of mislabelling, the peaks typically appear at the tails of the spectrum for configurations with large $\Delta\mu$ and $\theta = 180^\circ$.

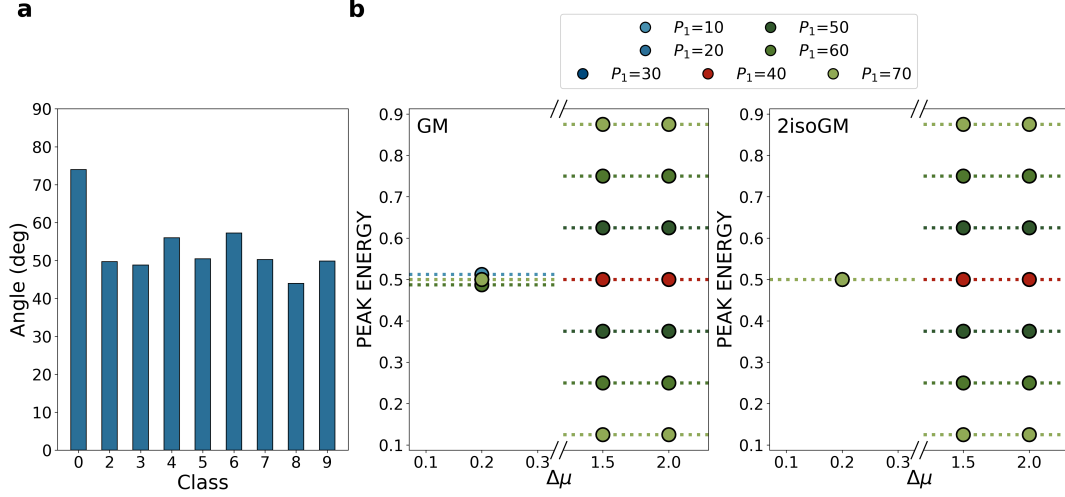


Figure 8.5: *DoS analysis for binary classification of Gaussian clones of MNIST data, investigating the impact of high-order statistics on DoS properties. (a) Angle between the mean vector of class 1 and the mean vectors of other classes in the MNIST dataset, as indicated on the x-axis. (b) Peak location of the DoS plotted against the inter-class separation distance $\Delta\mu$ for different values of class imbalance, as shown in the legend. Results are obtained from 9 different binary classification problems, comparing MNIST class 1 against each of the other MNIST classes. Red points indicate balanced classes, while blue and green points represent unbalanced classes, with lighter colours corresponding to greater imbalance. The two panels differ in the approximation used for the covariance matrix: Gaussian mixture (GM, left) and isotropic Gaussian mixture (2isoGM, right) clones.*

As the level of mislabelling L increases, the peaks move progressively toward the centre of the spectrum, converging to an energy of approximately 0.5. This behaviour reflects the increasing randomness of the classification task: with more mislabelled data, the network encounters a loss landscape where a large fraction of weight configurations yield similar levels of classification error, tending towards the limit of random classification. In this scenario, the DoS becomes dominated by configurations that provide no clear decision boundary between the classes. This transition aligns with the expectation that a heavily mislabelled dataset effectively loses its structure, rendering classification equivalent to chance.

This shift in peak position due to mislabelling mirrors the behaviour observed with class imbalance at low values of the angles between mean vectors (Figure 8.3.d), highlighting the complex interplay between data characteristics and the resulting DoS. The progression of the DoS toward a central energy point emphasizes the sensitivity of neural network performance to label noise and underscores the importance of ensuring dataset quality in real-world applications (Refinetti et al., 2023; Belrose et al., 2024).

Density of States in Classification Problems with Gaussian Clones of Real-World Datasets

An interesting technique to understand the relevance of input structure on learning is the construction of a hierarchy of clones that only retain some statistical properties

of the original real-world data.

In the light of this, we constructed a synthetic dataset where each class is a Gaussian clone of the original MNIST data, but where we can adjust the inter-class distances (see Supplementary Section I for further details). By doing so, we can reproduce the behaviour observed in the previous sections. This is true whether we employ the real correlation matrix (GM clone) or a diagonal covariance matrix (2ISOgm). In both cases, the DoS curves show properties that closely mirror those observed with the *synthetic* datasets discussed in Section 8.3.2, particularly how the peak location is influenced by class imbalance for small angles between the mean vectors (see Figure 8.5). The specific structure of the covariance matrices does not seem to play a role in determining the peak location, which is mostly influenced by the difference in the mean vectors.

8.4 Conclusions

In this chapter, a general approach to measuring the DoS in neural network learning problems has been introduced. Although enhanced sampling techniques are a standard tool in statistical mechanics and soft matter physics (Kim et al., 2006; Stelter et al., 2019; Kim et al., 2012; Tsai et al., 2007; Menichetti et al., 2021), their use in machine learning, and in neural networks in particular, has only recently begun to attract attention (Liu et al., 2023).

Here, for the first time to the best of our knowledge, one of these methods has been employed to thoroughly characterise the loss spectrum of simple neural architectures across a variety of learning tasks. A key aspect of this approach is that, unlike classic average-case analytical methods, it can be profitably applied to *finite-size single instances*, thus providing a complementary and promising perspective on a novel class of problems. Specifically, the WL algorithm has been used to clarify the impact of several factors on the DoS, with particular attention to the differential effects of input statistics, linear separability, and class imbalance. The latter point is especially relevant in view of the growing interest in the role of class imbalance in learning and in the implicit biases of deep neural networks (Francazi et al., 2024; Francazi et al., 2023). In particular, the analysis presented in this chapter addressed the effect of class imbalance on the entire spectrum of the training loss.

This first application of the methodology was restricted to fully connected networks with a single hidden layer. A natural next step is the extension to more complex and multilayered architectures, which would make it possible to investigate how the DoS is affected by weight sharing, as in convolutional layers, recurrent computation, as in sequence-processing architectures, and multiplicative interactions, as in attention-based models. While the extension of the technique to deeper networks is conceptually straightforward, its practical application raises an important issue of scalability (see Appendix F for a benchmark of convergence times). In the present context, applying the protocol to deep overparametrized networks

would entail a high computational cost, owing to the exploration of a configuration space that grows exponentially with network size. In this work, the original, “plain vanilla” implementation of the WL algorithm was adopted because of its simplicity and because it provides a suitable proof of concept for the present approach (F. Wang et al., 2001a; F. Wang et al., 2001b). The advantage of this choice in terms of clarity and simplicity is counterbalanced by limited efficiency, especially in deep overparametrized settings.

The scalability issue can, however, be addressed through several possible strategies. These include parallelisation over both training patterns and energy windows using GPU resources (Yin et al., 2012; Vogel et al., 2014; Farris et al., 2021), as well as more advanced variants of the WL algorithm based, for example, on multi-level sampling (Vogel et al., 2013), replica exchange (Vogel et al., 2014), and energy-biasing (Valentim et al., 2015). In soft and condensed matter physics, such techniques have been shown to significantly improve both sampling efficiency and scalability. In addition, (Liu et al., 2023) has demonstrated the advantages of optimised WL implementations in the context of neural networks. In that work, the method is used to explore the *input space* of a network, and improvements in efficiency and scalability are obtained through approximations such as entropy interpolation and gradient discretisation. Similar optimisations therefore appear promising for extending the scope of the present approach to larger and deeper neural architectures, and this direction is the subject of ongoing work.

The analysis in this chapter focused on networks with discrete synaptic weights. Nevertheless, WL techniques are known in soft and condensed matter physics to be applicable also in continuous spaces, as illustrated, for example, by their use in molecular simulations (Shell et al., 2002; Zhou et al., 2006; Yan et al., 2003). Extending the algorithm to continuous-weight neural networks does not alter the overall conceptual framework, but it does require more refined strategies for generating proposal moves in weight space. In contrast to the discrete case, however, such moves can exploit the differentiable nature of the loss landscape, so that strategies akin to stochastic gradient descent can be used to propose new weight configurations. Further support in this direction may also come from enhanced sampling techniques commonly employed in molecular simulation studies (Hénin et al., 2022).

Taken together, the possible improvements and extensions discussed above suggest that the present methodology may eventually become applicable to modern neural network models of practical relevance. For example, the approach could contribute to the study of the relation between weight-space and function-space descriptions, in line with recent developments on linearised models in Bayesian networks (Roy et al., 2024). This, in turn, could help clarify how weight configurations relate to functional outputs, complementing the information obtained from density-of-states analyses. More broadly, the WL approach may also provide useful indications for the selection of suitable architectures for a given problem, by

analysing the DoS together with the dynamics of learning algorithms as the weights traverse configurations with different entropy. Since gradient-based algorithms are known not to sample the loss landscape uniformly, but instead to display inductive biases (Farnia et al., 2020; Kalimeris et al., 2019; Rahaman et al., 2019) that can influence the entire learning dynamics (Refinetti et al., 2023; Belrose et al., 2024), studying optimisation dynamics from the perspective of the loss density appears to be a particularly interesting direction for future work.

Take-home message

The behaviour of a neural network is usually studied through trained solutions, but these represent only a small subset of all possible network configurations. In this chapter, Wang-Landau sampling was used to estimate the Density of States of neural networks, namely how many configurations correspond to each value of the classification error. This provided a global statistical description of the network configuration space and showed that its organisation depended jointly on the architecture and on the structure of the data. Learning can therefore be interpreted not only as the search for a good solution, but also as the exploration of a structured landscape of possible configurations.

9

Neural Network Dropout Space

9.1 Introduction

Modern neural networks have achieved remarkable empirical success, even as the principles through which learning, generalisation, and internal organisation emerge remain largely unknown. Among the numerous puzzling aspects of these systems that one might focus on in order to gather a hint of their inner workings, one is of particular scientific as well as technological prominence: the necessity for overparametrization. Indeed, modern neural networks are generally highly-overparametrized, in the sense that in order to achieve a high level of generalisation the system has to entail far more internal degrees of freedom than they are nominally required to encode the desired target input-output relation (Goodfellow et al., 2016; Zdeborová, 2020; Mehta et al., 2019; A. Engel et al., 2001).

This overparameterisation is widely regarded as a central ingredient in the optimisation dynamics and empirical success of neural networks; at the same time, though, it raises fundamental questions. Why does this apparent excess of parameters bear more benefit than damage? Are all these units necessary during training only or are in general required for the effective working of the network? How is task-relevant computation distributed across the hidden units? And, more precisely, given a trained hidden representation, can one identify specific neurons that are essential to preserve predictive performance and other that, instead, play a redundant or secondary role?

As anticipated, these questions bear directly on two distinct but closely related issues in deep learning: efficiency and interpretability. The first one pertains to the identification of those neurons that carry out most of the task-relevant computation; this is central to pruning and model compression, which seek to reduce memory usage and computational cost while preserving predictive performance (LeCun et al.,

1990; Hassibi et al., 1993; Han et al., 2016; Blalock et al., 2020; Hinton et al., 2015). Strategies for the reduction of network complexity, clearly while retaining the highest level of accuracy, are of key importance for deployment in resource-constrained settings, and they also have an impact on transferability, since a useful compressed representation should remain adaptable under fine-tuning on downstream tasks (Yosinski et al., 2014; Iofinova et al., 2022). Furthermore, model compression becomes extremely relevant when labelled data are scarce, because in that regime one would like to retain only the most robust and transferable features while avoiding unnecessary capacity.

The same motivations that fuel the quest for efficient model reduction approaches also play a role in the context of interpretability: in this context, one wonders whether specific hidden units have identifiable functional roles, how strongly information is localised or distributed, and whether dominant features can be disentangled within the learned representation (Bau et al., 2017; Morcos et al., 2018; Hooker et al., 2019). From this perspective, neuron selection is not only a compression strategy, but also a way to probe the network organisation that has emerged through training.

A large body of work has addressed these problems through pruning, sparsification, and regularisation methods. In these approaches, network parameters or even whole neurons are typically ranked according to criteria such as weight magnitude, local sensitivity, saliency, or estimated contribution to the loss, and the least important components are removed (LeCun et al., 1990; Hassibi et al., 1993; Han et al., 2016; Molchanov et al., 2017; Cheng et al., 2024). In parallel, dropout-based regularisation has shown that neural networks can remain effective even when substantial fractions of the hidden units are randomly suppressed during training, pointing to a considerable degree of redundancy in learned representations (Srivastava et al., 2014). These studies also indicate that redundancy is highly non-uniform: some subnetworks or subsets of units preserve function much better than others, as emphasised by works on winning tickets, neuron importance, and the behavioural effects of compression (Frankle et al., 2019; Morcos et al., 2018; Hooker et al., 2019; Blalock et al., 2020).

What remains much less clear, however, is whether these unequal contributions can be inferred directly from the internal statistical organisation of the hidden representation itself, rather than from supervised importance scores. This is a key issue with impactful fundamental as well as practical consequences: indeed, a positive answer to this question would imply that the efficiency of a trained network in performing a given task is discernible from statistical heuristics solely based on the hidden layer features, oblivious of external attributes associated with the task itself. The importance of a protocol to highlight these characteristics is hard to overestimate.

In this chapter, we address this question from an information-theoretic perspective. Rather than assigning neuron importance through their immediate effect on

the supervised loss, we treat the hidden layer as a configuration space generated by the network over the dataset, and define a reduced representation by retaining only a subset of hidden neurons. The central question is then whether the statistical structure of these hidden configurations, by itself, is sufficient to identify informative subsets. To this end, we use the **Mapping Entropy (ME)**, introduced in **Chapter 6**, which quantifies the information loss induced by a given **CG** choice. In the present setting, **CG** corresponds to a decimation of the hidden layer, and by minimising the **ME** one identifies the subset of neurons that best preserves the distinguishability of the hidden representations. In this way, the problem of neuron selection is recast as an unsupervised search for reduced hidden representations that retain the relevant configurational information.

This approach is interesting for two reasons. First, it provides an unsupervised criterion to probe the internal organisation of hidden layers, independently of gradient information or explicit supervised importance measures. Second, it allows us to test whether the information preserved by the representation is related to the actual functional contribution of the selected neurons, as assessed through the predictive performance of the reduced network. If such a correspondence holds, it would indicate that the geometry of the hidden representation already encodes which neurons contribute most to the task performed by the network.

The analysis is carried out on two complementary classes of systems. The first consists of **TS** networks, where the structure of the hidden representation is controlled by construction and the alignment between hidden units and teacher features can be monitored explicitly (A. Engel et al., 2001; Saad et al., 1995). This setting allows us to ask in an explicit way which neurons are selected by **ME** minimisation and how that selection evolves during training. The second is the **Non-Linear Gaussian Process (NLGP)**, which provides a synthetic but richer scenario in which hidden units naturally split into distinct functional classes (Ingrosso et al., 2022). Here the goal is to assess whether **ME** can detect such heterogeneity directly from the hidden representations. Finally, the functional significance of the selected subsets is evaluated through pruning experiments on **NLGP** and MNIST, by comparing the predictive performance of the corresponding reduced networks against random baselines (LeCun et al., 1998; Deng, 2012).

9.2 Materials and Methods

The hidden layer of a feed-forward neural network is taken as the microscopic representation of the system. A **CG** description is then defined by selecting a subset of $n_{cg} < K$ neurons from a hidden layer of width K . Accordingly, a decimation mapping $\mathcal{M}$ specifies which neurons are retained and therefore determines the reduced representation. In particular, given an input pattern $\mathbf{x}^\mu$, the network produces a hidden post-activation vector $\mathbf{h}^\mu = (h_1^\mu, \dots, h_K^\mu)$. Its components are binarized by taking their sign, yielding the binary hidden configuration

$$\boldsymbol{\phi}^\mu = (s_1^\mu, \dots, s_K^\mu), \quad s_i^\mu = \text{sign}(h_i^\mu) \in \{-1, +1\}. \quad (9.1)$$

The collection of such configurations over the dataset defines the ensemble of microscopic states considered in the analysis. For each $\boldsymbol{\phi}^\mu$, the mapping $\mathcal{M}$ induces a reduced configuration $\boldsymbol{\Phi}^\mu = \mathcal{M}(\boldsymbol{\phi}^\mu)$. At fixed reduced size n_{cg} , the optimal subset is identified by minimising the ME over the set $\mathcal{M}_{n_{cg}}$ of all mappings that retain exactly n_{cg} neurons:

$$\mathcal{M}^\star = \arg \min_{\mathcal{M} \in \mathcal{M}_{n_{cg}}} S_{\text{map}}(\mathcal{M}). \quad (9.2)$$

The formal definition of S_{map} and its information-theoretic interpretation were provided in Chapter 6. The optimisation is carried out with different strategies depending on the size of the hidden layer and, correspondingly, on the number of admissible mappings. When the mapping space is sufficiently small, the minimum is determined through an exhaustive exploration of all configurations. For larger hidden layers, where such a complete enumeration becomes computationally prohibitive, the minimum is instead approximated through a stochastic minimisation procedure based on simulated annealing.

9.2.1 Teacher-Student

As an initial controlled setting, we consider a one-hidden-layer network within the TS framework (Goldt et al., 2019; Goldt et al., 2020). The general TS construction was introduced in Chapter 6; here we specify the definitions and settings adopted in this work. Both teacher and student take inputs in $\mathbb{R}^N$ and have hidden layers of size M and $K > M$, respectively. This yields a controlled over-realised TS setting, in which the student has more hidden units than the teacher, while the teacher provides a natural reference structure (Watkin et al., 1993; A. Engel et al., 2001). Both networks are modelled as soft committee machines with fixed second-layer weights under the standard TS normalisation, and training acts only on the student first-layer weights (Saad et al., 1995; A. Engel et al., 2001).

Two instances of this setting are considered. The first is an analytically constructed replicated configuration, in which the student hidden units are organised into groups of replicas associated with the teacher units. The second consists of trained student networks obtained by online learning on teacher-generated examples. We first introduce the common TS architecture and notation, and then specify these two cases separately.

The teacher first-layer weights are taken to be an orthonormal set $\{\mathbf{B}_i\}_{i=1}^M$, with $\mathbf{B}_i \in \mathbb{R}^N$ and $\mathbf{B}_i \cdot \mathbf{B}_j = \delta_{ij}$. Accordingly, the student has $K = nM$ hidden neurons, indexed by pairs (i, α) , where $i \in \{1, \dots, M\}$ labels the associated teacher neuron and $\alpha \in \{1, \dots, n\}$ labels the replica within that group. Input patterns are sampled independently from the standard Gaussian distribution, $\mathbf{x}^\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. The teacher

and student outputs are defined by

$$y_T^\mu(\mathbf{x}) = \frac{1}{\sqrt{M}} \sum_{i=1}^M \phi(\mathbf{x}^\mu \cdot \mathbf{B}_i) = \frac{1}{\sqrt{M}} \sum_{i=1}^M t_i^\mu, \quad (9.3)$$

$$y_S^\mu(\mathbf{x}) = \frac{\sqrt{M}}{K} \sum_{i=1}^M \sum_{\alpha=1}^n \phi(\mathbf{x}^\mu \cdot \mathbf{J}_{i,\alpha}) = \frac{\sqrt{M}}{K} \sum_{i=1}^M \sum_{\alpha=1}^n s_{i,\alpha}^\mu, \quad (9.4)$$

where the activation function is

$$\phi(x) = \text{erf}\left(\frac{x}{\sqrt{2}}\right). \quad (9.5)$$

In the controlled setting, student first-layer weights are generated directly from the teacher weights through a tunable mismatch parameter $\eta \in [0, 1]$. Specifically,

$$\mathbf{J}_{i,\alpha} = \sqrt{1 - \eta^2} \mathbf{B}_i + \eta \mathbf{v}_{i,\alpha}, \quad (9.6)$$

where $\mathbf{v}_{i,\alpha}$ is a unit vector drawn uniformly in the subspace orthogonal to $\mathbf{B}_i$, so that $\mathbf{v}_{i,\alpha} \cdot \mathbf{B}_i = 0$. By construction, each student vector remains normalised and has overlap

$$\mathbf{J}_{i,\alpha} \cdot \mathbf{B}_j = \sqrt{1 - \eta^2} \delta_{ij}, \quad (9.7)$$

with the teacher weights. The parameter η therefore controls the degree of alignment: $\eta = 0$ corresponds to exact replication of the teacher representation, while $\eta = 1$ corresponds to a fully orthogonal representation within the complementary subspace.

Alongside the controlled construction, trained student networks were also analysed. The student was trained by an online learning procedure on examples generated by the teacher, with only the first-layer weights updated and the second-layer weights kept fixed at their prescribed normalised values. As noted above, training naturally drives the student towards the replicated organisation that is enforced analytically in the *in silico* construction. To apply the same (i, α) labelling to trained networks, each student neuron is assigned to the teacher class with which it has the largest overlap at the end of training. With this convention, an effective mismatch parameter is defined as

$$\eta_{i,\alpha}^{\text{eff}}(t) = \sqrt{1 - \left(\max_j \mathbf{J}_{i,\alpha}(t) \cdot \mathbf{B}_j \right)^2}, \quad (9.8)$$

and its network average as

$$\eta_{\text{eff}}(t) = \frac{1}{K} \sum_{i,\alpha} \eta_{i,\alpha}^{\text{eff}}(t). \quad (9.9)$$

The quantity $\eta_{\text{eff}}(t)$ provides a time-dependent measure of the student-teacher alignment, enabling a direct comparison between trained networks and the controlled *in silico* reference.

For each value of η (or η_{eff}), the ME analysis was performed over subsets of student neurons. Since the replicated construction associates each student neuron with a well-defined teacher class, the selected mappings can be characterised not only by their ME value but also by their composition. For a mapping $\mathcal{M}$, let $n_i(\mathcal{M})$ denote the number of selected student neurons associated with teacher neuron i , and define

$$p_i(\mathcal{M}) = \frac{n_i(\mathcal{M})}{|\mathcal{M}|}, \quad i = \{1, \dots, M\}, \quad (9.10)$$

with $\sum_{i=1}^M p_i(\mathcal{M}) = 1$. The balance of the mapping is then quantified by

$$\Delta(\mathcal{M}) = \frac{1 - \max_i p_i(\mathcal{M})}{1 - 1/M}. \quad (9.11)$$

By construction, $\Delta = 0$ when all selected neurons belong to the same class, corresponding to a maximally redundant mapping, whereas $\Delta = 1$ when the selected subset is perfectly balanced across all M classes. The quantity Δ was used throughout the analysis as a complementary descriptor of the structure of the selected mappings, and its definition applies without modification to both the *in silico* and trained settings.

We investigated two TS configurations, both with input dimension $N = 100$. In the first, the teacher and student hidden layers have sizes $M = 2$ and $K = 10$, respectively, and the analysis is based on 3×10^4 examples; this system is small enough to allow an exhaustive exploration of the mapping space. In the second, $M = 5$ and $K = 25$, and the analysis is based on 10^5 examples; in this case, the larger mapping space makes it necessary to minimise the ME by simulated annealing rather than by exhaustive search. Apart from this difference in numerical treatment, the model construction and the definitions of η , η_{eff} , and Δ are identical in the two settings.

9.2.2 Non-Linear Gaussian Process

As a second model, we considered a synthetic binary classification task based on a Non-Linear Gaussian Process (NLGP) (Ingrosso et al., 2022). Input patterns are N -dimensional vectors, where each component x_i , with $i = 1, \dots, N$, plays the role of a pixel intensity. Two classes are generated, corresponding to Gaussian fields with correlation lengths ξ^+ and ξ^- , respectively. The patterns belonging to each class are generated as

$$\mathbf{x}^{\xi^\pm} = \frac{\psi(\gamma \mathbf{z}^{\xi^\pm})}{Z(\gamma)}, \quad (9.12)$$

where $\mathbf{z}^{\xi^\pm} \in \mathbb{R}^N$ is a zero-mean Gaussian vector with covariance

$$C_{ij}^{\xi^\pm} = \langle z_i^{\xi^\pm} z_j^{\xi^\pm} \rangle = \exp\left(-\frac{|i-j|^2}{(\xi^\pm)^2}\right), \quad (9.13)$$

with periodic boundary conditions imposed on the input coordinate index. The two classes are therefore distinguished by their respective correlation lengths ξ^+ and ξ^- . The non-linear function $\psi(z) = \text{erf}(z/\sqrt{2})$ introduces non-Gaussian statistics controlled by the parameter γ , while the normalisation factor $Z(\gamma)$ is chosen such that $\text{Var}(\mathbf{x}^{\xi^\pm}) = 1$. Throughout the analysis, the values $N = 50$, $\xi^+ = 7.5$, $\xi^- = 2$, and $\gamma = 15$ were employed. Both the training and test sets contain a total of $P = \alpha N$ patterns, with $\alpha = 300$, equally divided between the two classes. The labels assigned to the two classes are $y = +1$ and $y = -1$, respectively.

The network is a one-hidden-layer perceptron with N input units, $K = 30$ hidden neurons, and a single linear output unit. The hidden-layer activation function is

$$\phi(x) = \text{erf}\left(\frac{x}{\sqrt{2}}\right), \quad (9.14)$$

and the output weight vector is fixed to $w_i^{\text{out}} = 1/\sqrt{K}$ for all i , with zero output bias; only the first-layer weights and biases are trained. Training was performed by stochastic gradient descent on the mean squared error loss with L_2 regularisation coefficient 0.1, learning rate 0.1, and batch size 300. Networks were trained for up to 3×10^4 epochs.

For the analysis of trained networks, hidden neurons were partitioned into two empirical groups according to the structure of their incoming weight vectors. Neurons with weight vectors concentrated on a restricted subset of input coordinates were classified as localised, whereas neurons with extended alternating-sign profiles were classified as oscillatory (Ingrosso et al., 2022). The two groups were quantitatively identified by means of the inverse participation ratio (IPR), defined for each hidden neuron i as

$$\text{IPR}(\mathbf{w}_i) = \frac{\sum_{j=1}^N w_{ij}^4}{\left(\sum_{j=1}^N w_{ij}^2\right)^2}, \quad (9.15)$$

which ranges between $1/N$, attained when all weights are equal in magnitude, and 1, attained when only a single weight is non-zero. Localised neurons therefore exhibit systematically larger IPR values than oscillatory ones. The two groups also develop distinct bias values during training, with oscillatory neurons associated with positive bias and localised neurons with negative one, thus providing a complementary indicator for their identification and tracking.

The ME analysis was performed at multiple checkpoints during training. At each epoch, the hidden representations were computed by passing all training patterns through the network and binarising the continuous activations via the sign

function. The ME minimisation was then carried out independently at each checkpoint and for each value of the number of neurons retained n_{cg} , using multiple independent stochastic minimisation runs. Additionally, the composition of the selected mappings can be also characterised in terms of the fraction of localised neurons $f_{loc}(\mathcal{M})$ among the n_{cg} retained units,

$$f_{loc}(\mathcal{M}) = \frac{n_{loc}(\mathcal{M})}{|\mathcal{M}|}. \quad (9.16)$$

9.2.3 Wang-Landau Sampling of the Dropout Space

In addition to the direct minimisation of the ME, the structure of the mapping space of the NLGP system was characterised via WL sampling in the $1/t$ formulation (F. Wang et al., 2001b; Belardinelli et al., 2007b). A more detailed theoretical and methodological discussion of the WL approach is provided in Chapter 6, to which the reader is referred. The present section therefore focuses only on the aspects of the procedure that are specific to the case study considered here. Within this framework, the ME, $S_{map}(\mathcal{M})$, was treated as the effective energy associated with a mapping $\mathcal{M}$. Sampling was terminated when the refinement parameter reached 10^{-5} .

At fixed retained neuron size n_{cg} and fixed number of localised neurons n_{loc} , the mapping space is defined as the set of all subsets of n_{cg} hidden neurons containing exactly n_{loc} localised units or, analogously, $n_{cg} - n_{loc}$ oscillatory units.

The WL analysis was performed at fixed $n_{cg} = 15$. This choice is particularly relevant for the present system, as the hidden layer consists of 30 neurons, evenly divided into 15 localised and 15 oscillatory units, making $n_{cg} = 15$ a natural CG size for the analysis. For each admissible value of n_{loc} , an independent WL simulation was carried out on the corresponding constrained mapping space. Trial moves were defined by exchanging one (or more) retained neuron(s) with one (or more) discarded neuron(s) belonging to the same class, thereby preserving both n_{cg} and n_{loc} throughout the sampling dynamics. Each simulation was continued until the refinement parameter reached 10^{-5} .

The quantity of interest is the joint Density of States (DoS)

$$g(S_{map}, n_{loc}), \quad (9.17)$$

or, equivalently, its representation in terms of the fraction of localised neurons $f_{loc} = \frac{n_{loc}}{n_{cg}}$. The joint DoS was reconstructed by combining the independently sampled spectra obtained at fixed n_{loc} . This reconstruction is possible because each simulation explores the full range of effective energies accessible under the corresponding compositional constraint and because, for fixed n_{cg} and n_{loc} , the total number of admissible mappings is known exactly. These combinatorial counts provide the appropriate normalisation for each restricted DoS, thereby allowing the different

contributions to be assembled into the full two-dimensional DoS. The same analysis was performed at different training epochs in order to characterise the evolution of the mapping space structure during learning.

9.2.4 Performance Evaluation of Reduced Networks

To test whether the mappings selected by ME minimisation also identify effective reduced representations, we construct pruned networks by retaining only the hidden neurons selected by the mapping. For each n_{cg} value, the performance of the selected subset is compared with that obtained from random subsets of the same size. In the NLGP case, pruning is applied to the hidden layer of the trained classifier. After selecting the retained neurons, the output bias is re-optimised while keeping the hidden-layer parameters fixed.

As a second classification benchmark, we consider a feed-forward neural network with a single hidden layer of $K = 30$ neurons and erf activation function. The task is binary classification on the MNIST dataset (LeCun et al., 1998; Deng, 2012), restricted to digits 1 and 7. To make the problem transnationally invariant and more challenging, the dataset is augmented by random translations of the input images. The network is trained by stochastic gradient descent, with only the first-layer parameters optimised. Reduced networks are then obtained by pruning the hidden layer according to the selected mapping, and their test accuracy is compared with that obtained from random subsets of the same size.

9.3 Results

9.3.1 Structure of the Mappings Selected by Mapping Entropy Minimisation

In this section, we analyse the properties of the subsets selected by ME minimisation in controlled settings where the organisation of the hidden representation is already understood. To this end, we consider two complementary scenarios: the overparameterised TS regression model, where hidden units can be interpreted through their alignment with the teacher, and the NLGP classification problem, where neurons organise into distinct functional classes. In both cases, this well-defined structure, known *a priori*, makes it possible to interpret the subsets selected by ME.

Teacher-Student system

We begin with the controlled TS setting, where the structure of the student hidden layer is fixed by construction and therefore provides a direct reference for interpreting the subsets selected by ME minimisation. The relevant observable is the balance index Δ , shown in Figure 9.1e,f as a function of the mismatch parameter η for different numbers of retained neurons n_{cg} . By construction, $\Delta = 1$ identifies a mapping

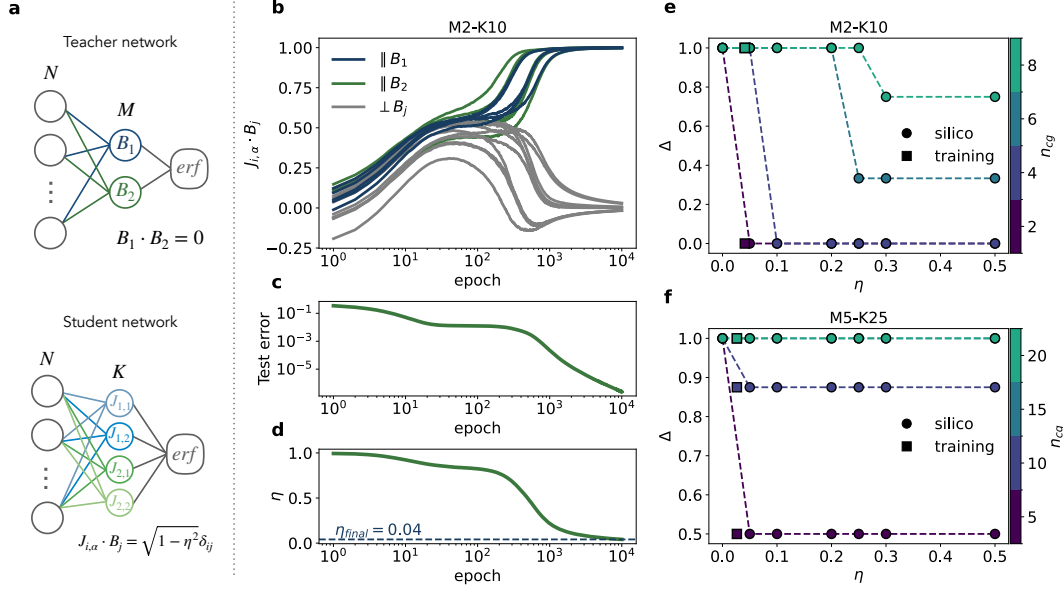


Figure 9.1: *TS system and ME analysis of hidden-layer representations.* (a) Schematic representation of the teacher and student networks. The teacher has M hidden units with orthogonal first-layer weights $\mathbf{B}_i$, while the student has $K > M$ hidden units labelled by (i, α) . In the controlled construction, the student weights satisfy $\mathbf{J}_{i,\alpha} \cdot \mathbf{B}_j = \sqrt{1 - \eta^2} \delta_{ij}$. (b) Training dynamics of the overlaps $\mathbf{J}_{i,\alpha} \cdot \mathbf{B}_j$ in the M2-K10 case, shown as a function of the training epoch. Blue and green curves denote the overlaps with $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively, while grey curves denote the components orthogonal to the teacher directions. (c) Generalization error as a function of the training epoch for the same system. (d) Effective mismatch parameter η_{eff} as a function of the training epoch. The dashed horizontal line indicates the final value $\eta_{\text{final}} = 0.04$. (e, f) Balance Δ of the ME-minimising mapping as a function of the mismatch parameter η , for the M2-K10 and M5-K25 systems, respectively. Different colours identify different values of the retained CG dimensionality n_{cg} , as indicated by the colour bar. Circular markers denote the controlled in silico construction, and square markers denote the trained networks, plotted at the corresponding effective mismatch value η_{eff} . Note the different scale of the Δ axis in the two plots.

that distributes the retained neurons as evenly as possible across teacher groups, whereas smaller values correspond to progressively more concentrated selections. Values close to $\Delta = 1$ are therefore desirable, because they indicate that the selected subset preserves the full teacher structure rather than over-representing only a restricted part of it.

We start from the exact replicated limit, $\eta = 0$, which provides the natural reference case for the discussion. Here, the ME-minimising mapping is fully balanced ($\Delta = 1$) at the smallest number of neurons compatible with the teacher structure: $n_{\text{cg}} = 2$ for the M2-K10 system and $n_{\text{cg}} = 5$ for the M5-K25 system (Figure 9.1.e,f). Thus, when the student exactly replicates the teacher representation, the entropy minimum coincides with the smallest subset that still covers all teacher modes. This is a non-trivial result, because the ME is computed solely from the statistics of the hidden configurations and has no explicit access to the teacher labels; nevertheless, in the replicated limit it selects one representative degree of freedom for each teacher mode.

As soon as the mismatch becomes finite, deviations from this correspondence

begin to emerge. In both systems, the curves in Figure 9.1.e,f show that the smallest number of neurons for which the selected mapping is balanced increases monotonically with η . In the M2-K10 case, the mapping selected at the minimal size $n_{cg} = 2$ is balanced only at $\eta = 0$, while already at $\eta \approx 0.1$ it has collapsed to $\Delta = 0$. By contrast, the first n_{cg} value that remains balanced over an extended mismatch range is $n_{cg} = 4$, which keeps $\Delta = 1$ up to $\eta \approx 0.2$. A similar trend is visible in the M5-K25 case, but on a broader scale: the minimal balanced mapping occurs at $n_{cg} = 5$ only at $\eta = 0$, while for finite mismatch the balanced solution is shifted to substantially larger numbers of retained neurons. For instance, curves with $n_{cg} \approx 10$ remain at $\Delta \approx 0.875$, showing that they still over-represent some teacher groups, whereas the first fully balanced solution appears only for $n_{cg} = 15$.

This systematic shift has a clear interpretation. The ME does not probe functional equivalence with the teacher, but the statistical distinguishability of the hidden units. At $\eta = 0$, different replicas associated with the same teacher unit are statistically indistinguishable, so one representative per group is sufficient. At finite η , each student neuron acquires an additional orthogonal component and replicas are no longer equivalent at the level of hidden-layer statistics. As the variability of the hidden representation increases, the overall minimum of this profile shifts towards larger n_{cg} values, because more neurons must be retained to resolve the additional variability. In this sense, the increase of the optimal n_{cg} is not simply a failure to recover the teacher-minimal representation; rather, it quantifies how far the hidden representation is from the exact replicated limit.

It is then natural to ask whether the same scenario survives in trained networks, where the variability is not imposed homogeneously by construction but generated by the learning dynamics. Panels Figure 9.1.b–d show that training indeed drives the student towards the teacher representation. In the M2-K10 example, the overlaps with the two teacher directions increase from values close to zero at the beginning of training to values close to one at late times, while the orthogonal components are progressively suppressed (Figure 9.1.b). Over the same interval, the generalisation error decreases by several orders of magnitude (Figure 9.1.c), and the effective mismatch η_{eff} drops from approximately 1 to a final value $\eta_{final} \approx 0.04$ (Figure 9.1.d). The trained student therefore approaches, but does not exactly reach, the replicated limit.

The ME analysis of the trained networks is fully consistent with the controlled construction. In Figure 9.1.e,f, the square markers corresponding to the trained networks fall on the same branches defined by the *in silico* curves when plotted at the corresponding value of η_{eff} . Quantitatively, the trained networks no longer recover a balanced mapping at the strictly minimal size $n_{cg} = M$: for M2-K10 the smallest balanced mapping is found at $n_{cg} = 4$, and for M5-K25 at $n_{cg} = 15$. These are precisely the same n_{cg} values selected by the controlled curves at mismatch values of order $\eta \approx 0.04$, showing that the trained networks behave as weakly mismatched replicas of the teacher representation.

Taken together, these results identify a clear selection principle in the **TS** setting. In the exact replicated limit, **ME** minimisation recovers the minimal teacher-consistent hidden representation. As soon as replicas are no longer statistically equivalent, the n_{cg} value for which $\Delta = 1$ increases, and the amount by which it increases provides a direct measure of the residual variability present in the hidden layer. The agreement between the analytically controlled construction and the trained networks shows that this interpretation remains valid beyond the idealised setting and continues to hold in the presence of heterogeneous, learning-induced correlations.

NLGP System

A richer and less constrained setting is provided by the **NLGP** classification problem, where the hidden representation is structured, but not through explicit replication. The full set of results is shown in **Figure 9.2**. As already known from previous studies (**Ingrosso et al., 2022**), after training the hidden neurons separate into two classes, illustrated in **Figure 9.2.b**: localised neurons, whose weights are concentrated on a restricted subset of input coordinates, and oscillatory neurons, whose weights display an extended alternating-sign profile. Here this prior structural information is used as a reference frame to analyse the mappings selected by the **ME** minimisation.

Before turning to the **ME** analysis, it is important to establish that the two neuron classes correspond to distinct yet individually informative representations of the classification task. This is shown in **Figure 9.2.c**, where the network output is evaluated after retaining only one class of hidden neurons at a time. In both cases, the outputs associated with the two input classes remain clearly separated, showing that either population alone is sufficient to preserve discriminative information. The difference lies in the form of the representation: localised neurons and oscillatory neurons induce distinct output distributions, reflecting two qualitatively different internal encodings of the same classification rule. **Figure 9.2.c** therefore shows that the two populations should not be interpreted as playing complementary roles that become meaningful only when combined; rather, they define alternative representational modes, each of which can by itself support the separation of the input classes.

The **ME** analysis shows that the optimal mappings do not combine these two representations arbitrarily. **Figure 9.2.d** reports the average fraction of localised neurons in the optimal mapping as a function of n_{cg} , for different training epochs. At early times, the selected mappings are almost entirely localised over the whole range of retained neuron subset sizes: at epoch 100, the localised fraction is essentially unity for all n_{cg} shown. The corresponding selection-probability map in the right panel confirms this picture: at epoch 100, the optimisation repeatedly selects the same band of localised neurons with $p_{in} \approx 1$, whereas oscillatory neurons are

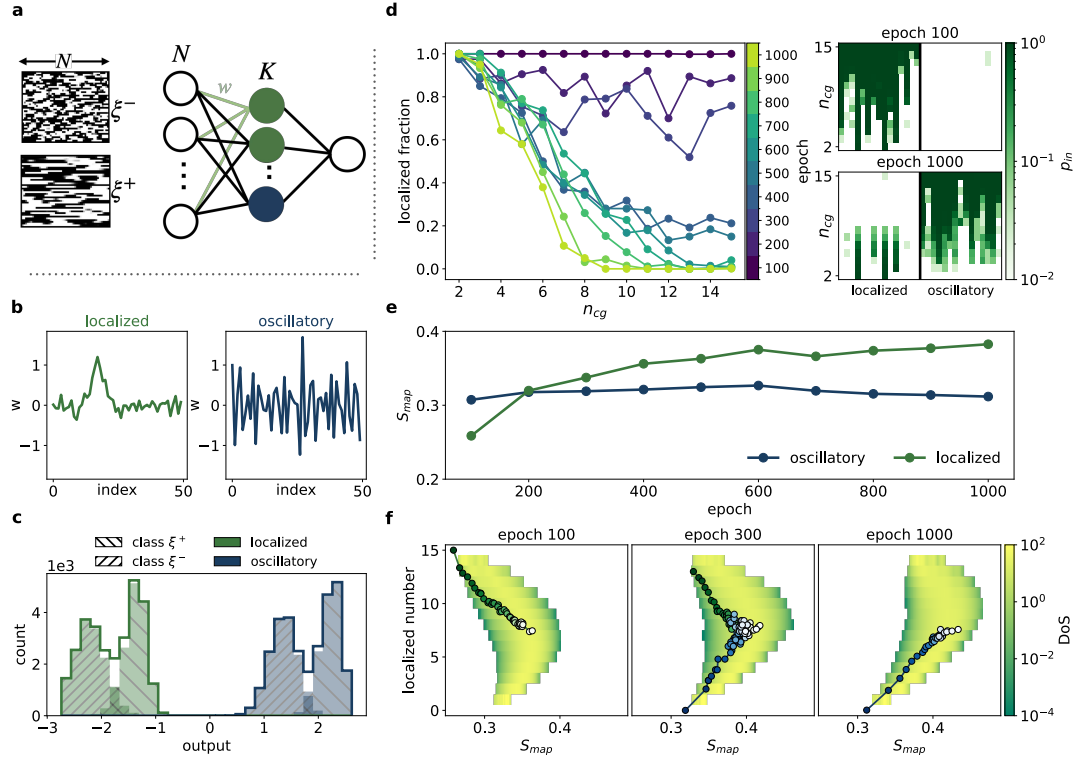


Figure 9.2: NLGP system and ME analysis of hidden-layer representations. (a) Schematic of the classification task and network architecture. Input patterns from the two classes, characterised by different correlation lengths, are shown alongside the one-hidden-layer network with $K = 30$ neurons. (b) Representative examples of hidden neurons after training: a localised neuron (left), with weights concentrated on a subset of input coordinates, and an oscillatory neuron (right), with extended alternating-sign structure. (c) Distribution of the network output evaluated using only localised neurons (green) or only oscillatory neurons (blue), compared across the two classes (hatched histograms). (d) Left: average fraction of localised neurons in the ME-minimising mapping as a function of the number of retained neurons n_{cg} , for different training epochs (colour-coded). Right: selection probability p_{in} of each neuron, shown as a function of n_{cg} , for two representative epochs (100 and 1000). Neurons are ordered according to their bias value. (e) ME S_{map} evaluated for mappings composed exclusively of localised neurons (green) or oscillatory neurons (blue), as a function of training epoch. (f) DoS of the mapping space as a function of S_{map} and the number of localised neurons, for different training epochs. Overlaid points indicate average optimisation trajectories.

almost never included.

As training proceeds, this preference changes in a strongly size-dependent way. Around epoch 300, the average localised fraction starts to decrease for intermediate and large n_{cg} values, signalling the appearance of a competing oscillatory solution. By late training, the transition is sharp: at epoch 1000, the optimal mappings are still almost fully localised for very small subsets, but the localised fraction drops below $1/2$ already around $n_{cg} \simeq 6-7$, and becomes essentially zero for $n_{cg} \gtrsim 10$. The heat map at epoch 1000 shows that this crossover does not arise from a gradual mixing of the two populations. Rather, for small n_{cg} the selected neurons come almost exclusively from the localised group, whereas for larger n_{cg} the selected mappings are almost entirely oscillatory.

This interpretation is corroborated by the direct comparison shown in Fig-

ure 9.2.e at $n_{cg} = 15$, where the retained subset coincides with the full set of oscillatory neurons or, alternatively, with the full set of localised neurons. At epoch 100, purely localised mappings have a substantially lower ME than purely oscillatory ones, with $S_{map} \approx 0.26$ for the localised case against $S_{map} \approx 0.31$ for the oscillatory one. Around epoch 200, the two values become comparable, $S_{map} \approx 0.32$, marking the point at which the two representational modes compete on equal footing. At later epochs the ordering is reversed: the value of the ME of selections containing exclusively oscillatory neurons remains approximately flat in the interval 0.31–0.33, whereas the one including only localised neurons increases steadily up to ≈ 0.38 by epoch 1000. The growing gap quantitatively explains why, at late times, oscillatory mappings dominate the optimal set at large n_{cg} .

The WL reconstruction of the full mapping space provides a more global view of this transition. In Figure 9.2.f, the DoS is shown as a function of S_{map} and the number n_{loc} of localised neurons retained in mappings of fixed size $n_{cg} = 15$. A first important point is that the ME minima are generally not found in mixed mappings containing substantial contributions from both functional classes; rather, they lie close to the edges of the space, where the retained subset is dominated by one class or the other. At epoch 100, the low- S_{map} basin is located at large n_{loc} , at the fully localised edge, and the optimisation trajectories converge to this region. At epoch 300, the landscape becomes effectively bimodal: two competing basins are visible, and the optimisation trajectories split between them. By epoch 1000, the low- S_{map} basin has moved to $n_{loc} = 0$, *i.e.*, to purely oscillatory mappings, while the localised side remains only as a higher- S_{map} sector of the mapping space. The change observed in Figure 9.2.d is therefore not a finite-sampling artefact of the annealing dynamics, but reflects a true restructuring of the ME landscape.

Overall, the NLGP results show that ME minimisation does not simply favour sparse or mixed subsets. Instead, it selects coherent mappings associated with one of the two pre-existing representational classes, and the preferred class depends both on the stage of training and on the number of neuron retained. At early epochs, localised neurons provide the most informative reduced description. During training, an oscillatory representation progressively becomes competitive and eventually dominates for sufficiently large n_{cg} . The ME therefore acts as a probe of the internal organisation of the hidden layer, resolving not only which neurons class should be retained, but also which representational mode furnishes the statistically sharpest CG description.

9.3.2 Performance of the Reduced Network

Having characterised the structure of the mappings selected by ME minimisation in controlled settings, we now turn to their functional validation. The goal is to determine whether these subsets, beyond exhibiting a clear and non-trivial organisation in the benchmark cases discussed above, also preserve the predictive performance

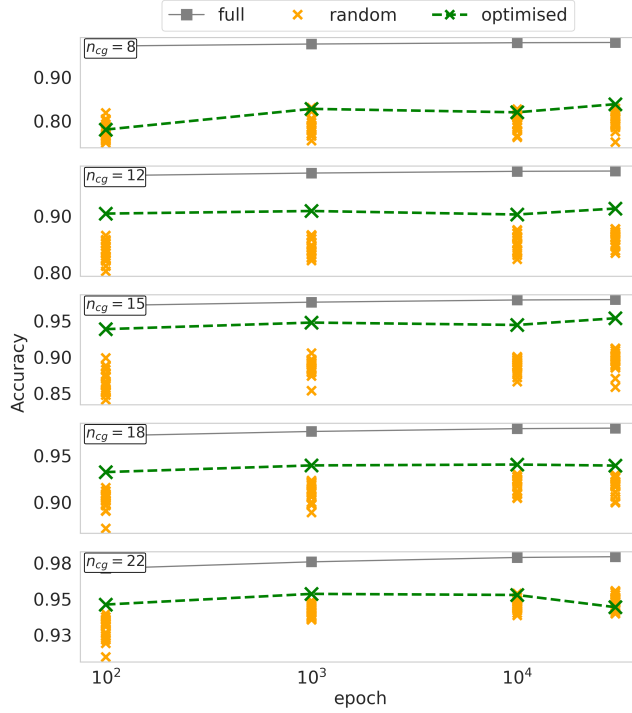


Figure 9.3: *NLGP* reduced-network performance at different training epochs. Classification accuracy as a function of the number of retained neurons n_{cg} . For each n_{cg} , the green markers denote the accuracy of the reduced network obtained from the *ME*-selected subset, the orange markers the accuracies obtained from random subsets of the same size, and the dashed grey line the accuracy of the full network. After pruning, the output bias is re-optimised while keeping the hidden-layer parameters fixed. Note that the y axes do not start from zero.

of the original network after pruning. Because the *ME* criterion is defined solely from the statistics of the hidden activation patterns, with no direct access to the task or to the output labels, this constitutes a stringent test of its effectiveness as an unsupervised pruning criterion.

We begin with the *NLGP* classifier, where the performance of reduced networks can be monitored at different stages of training and for several values of retained neurons n_{cg} (Figure 9.3). For each n_{cg} value, the accuracy obtained from the *ME*-selected subset is compared with that of random subsets of the same cardinality, as well as with the accuracy of the full network. After pruning, the output bias is re-optimised while keeping the hidden-layer parameters fixed. This is a minimal readjustment, introduced to compensate for the shift of the output distribution induced by the selected hidden subset. Its use is motivated looking at Figure 9.2.c: localised and oscillatory populations define distinct, yet individually viable, representations of the classification rule, but they generally induce different offsets in the output distribution. As a consequence, pruning changes the effective decision threshold of the classifier. If the zero threshold used for the full network were kept unchanged, all reduced selections would yield an accuracy close to 0.5, with one of the two classes being almost entirely misclassified. Re-optimising the output bias therefore removes this trivial source of performance loss and allows one to compare different subsets on the basis of the information retained in the hidden representation.

A clear trend emerges. For all epochs considered, the *ME*-selected subsets typically achieve an accuracy above the bulk of the random distribution, showing that the mappings identified through the activation statistics also retain a compara-

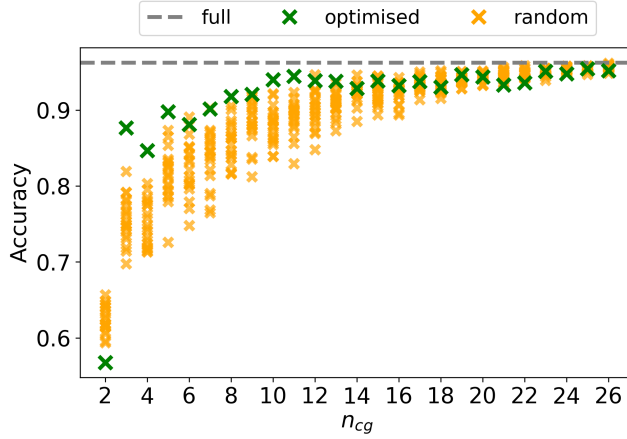


Figure 9.4: Reduced-network performance for the MNIST binary classification task (1 vs 7) with data augmentation by random translations. Classification accuracy as a function of the number of retained neurons n_{cg} . For each n_{cg} , the green markers denote the accuracy of the reduced network obtained from the ME-selected subset, the orange markers the accuracies obtained from random subsets of the same size, and the dashed grey line the accuracy of the full network. Note that the y axes do not start from zero.

tively informative representation for the task. This advantage is most visible in the strongly compressed regime. For instance, at $n_{cg} = 8$ and $n_{cg} = 12$, the optimised subset lies systematically above the typical random choice at all epochs. The same behaviour persists at $n_{cg} = 15$ and 18 , although the separation becomes progressively smaller as the number of retained neurons increases. By contrast, at $n_{cg} = 22$ the two procedures become nearly indistinguishable, as expected when most of the hidden layer is already retained. Overall, the NLGP results indicate that ME minimisation provides a non-trivial and functionally meaningful guide for pruning, especially when the compression is sufficiently strong (i.e. when up to half of the hidden layer neurons are culled) that the specific choice of subset matters.

The same analysis can then be extended to a more realistic classification problem, namely the binary MNIST task with data augmentation by random translations, using a network with $K = 30$ hidden-neurons. In this case, shown in Figure 9.4, we report the reduced-network performance only at the end of training, since the behaviour is qualitatively stable across epochs. Also here the accuracy obtained from the ME-selected subset is compared with the distribution generated by random subsets of equal size. The overall picture is consistent with the NLGP benchmark, but with a clearer dependence on the number of retained neurons. Excluding the extreme case $n_{cg} = 2$, where all reduced networks perform poorly, the ME-selected mappings systematically lie in the upper part of the random distribution for small and intermediate n_{cg} values. In particular, for $3 \lesssim n_{cg} \lesssim 12$, the selected subset typically yields an accuracy above the bulk of the random configurations and often close to their upper tail. This indicates that, in the strongly pruned regime, the ME criterion is able to identify hidden subsets that preserve most effectively the task-relevant information. For larger n_{cg} values, however, the distinction progressively weakens: the random and optimised subsets become compatible within the observed spread, and for the largest n_{cg} values random pruning can even perform comparably, or slightly better. This is consistent with the fact that, when the reduction is mild, the space of admissible subsets becomes much less heterogeneous and the benefit of a structured selection is correspondingly reduced.

Taken together, these results show that the mappings selected by **ME** minimisation are not only structurally meaningful, but also functionally relevant. The criterion does not identify the optimal reduced network in an absolute sense, nor should this be expected, since it is blind to the specific task and relies only on the hidden activation statistics. Nevertheless, in a broad range of regimes it provides an effective proxy for pruning. Its advantage is most evident when the compression is sufficiently strong that the choice of the retained subset is non-trivial, while it naturally diminishes as n_{cg} approaches the size of the full hidden layer.

9.4 Conclusions

In this chapter, we have investigated the dropout space of neural networks through the lens of **ME**, with the aim of identifying reduced representations that are both structurally meaningful and functionally effective. By framing neuron selection as a **CG** problem, we have extended the information-theoretic framework developed in the previous chapters to the domain of learning systems.

The analysis has revealed that **ME** minimisation selects highly non-trivial subsets of hidden neurons, whose properties reflect the internal organisation of the network. In controlled **TS** settings, the selected neurons are strongly aligned with the informative directions defined by the teacher, indicating that the **ME** is able to recover the relevant features of the representation without explicit supervision. In more complex scenarios, such as the **NLGP** system, the selection highlights functional heterogeneity within the hidden layer, systematically favouring neurons that contribute most to the variability of the representation.

Beyond this structural insight, a central result of the chapter is the demonstration that **ME**-based selection has direct functional implications. Reduced networks obtained by retaining the neurons selected through **ME** minimisation consistently outperform those constructed from random subsets of the same size. This establishes a concrete link between information-theoretic optimality and predictive performance, showing that the preservation of configurational distinguishability translates into improved generalisation capabilities.

These findings suggest that the **ME** captures structural information that is intrinsic to the data representation, rather than being tied to a specific task or loss function. In this sense, **ME** provides a principled pruning criterion that is complementary to standard approaches based on weight magnitude, sensitivity analysis, or gradient-based importance measures (LeCun et al., 1990; Molchanov et al., 2017). Unlike these methods, which rely on supervised signals, the **ME**-based approach operates directly on the statistical structure of the hidden configurations, making it applicable in a broader range of settings.

From a statistical physics perspective, the results reinforce the interpretation of neural networks as complex systems whose behaviour is governed by the geometry of their configurational space. The dropout space explored in this chapter can be

seen as a **CG** representation of this geometry, where different subsets of neurons correspond to different projections of the underlying landscape. The fact that **ME** identifies subsets that are both structurally interpretable and functionally optimal indicates that relevant degrees of freedom are encoded in the organisation of this space and can be extracted through information-theoretic principles.

Several directions emerge from this work. On the methodological side, it would be natural to extend the analysis to deeper architectures and to consider more general forms of **CG** beyond simple neuron decimation. On the theoretical side, a deeper understanding of the relationship between **ME** and generalisation remains an open question, particularly in the presence of strong correlations and non-linear interactions. Finally, from an applied perspective, the integration of **ME**-based criteria into practical pruning and compression pipelines could provide a novel route towards more efficient and interpretable approaches.

Take-home message

Removing hidden neurons from a trained neural network defines a space of dropout configurations, each corresponding to a reduced architecture with a different level of retained information. In this chapter, **Mapping Entropy** was used to rank these reduced representations and to identify subnetworks that minimised the information loss induced by pruning. The results showed that **Mapping Entropy**-selected subnetworks could preserve predictive performance while retaining only part of the original network, indicating that over-parameterised architectures contain redundancy, but also that their informative degrees of freedom are structured and detectable through information-theoretic criteria.

10

Conclusions

This thesis has investigated how meaningful information can emerge from low-resolution descriptions of high-dimensional systems. The central perspective throughout this work has been that complex systems are most naturally understood through the structure of their configurational space. The problem addressed has therefore been not only how reduced representations of such spaces can be constructed, but also how their organisation depends on the adopted resolution level and on the selected properties of the system under study. This work has approached these questions at two levels: a general theoretical perspective and the study of specific complex systems.

The more theoretical results have been presented in [Chapter 3](#) and [Chapter 7](#). In particular, [Chapter 3](#) showed that the [Resolution-Relevance \(RR\)](#) framework, based on two information-theoretic quantities that measure how finely a system is described and how much of that description reflects meaningful structure in the data, can provide a consistent unsupervised indication of representation size. Comparison with supervised benchmarks showed that the unsupervised optimum progressively converges towards the representation that best approximates the underlying generative probability distribution. This result supports the interpretation of [RR](#) as an unsupervised criterion for identifying a class of low-resolution representations that are maximally informative about the unknown generative process underlying the data.

[Chapter 7](#), in turn, addressed the complementary question of how the information loss induced by a reduced representation can be related to the interaction and correlation structure of the system. This loss is generally quantified by [Mapping Entropy \(ME\)](#), which measures how much information is discarded when passing from a detailed description of a system to a reduced one. [Chapter 7](#) presented the derivation of a multi-body decomposition of [ME](#), which showed that this loss can be resolved into one-body, two-body, and higher-order contributions. This multi-body structure leads to a more interpretable form of [ME](#), while also providing a natural basis for low-order approximations when the full quantity becomes difficult to eval-

uate or optimise.

The general framework outlined above has found concrete applications in the system-specific analyses, where the search for informative reduced representations was pursued in two very different domains: proteins, which are molecular systems exploring the configurational space of their constituent atoms; and learning networks, which are artificial adaptive systems defined in a high-dimensional parameter spaces. Despite the difference in physical nature and function, both display key features of complex systems, including many interacting components, high dimensionality, and emergent collective behaviour.

For proteins, [Chapter 4](#) and [Chapter 5](#) examined two complementary aspects of informative low-resolution description. [Chapter 4](#) focused on the [Coarse-Graining \(CG\)](#) of the structure space explored by [Molecular Dynamics](#) trajectories, and showed that the most informative procedures to cluster the sampled conformations are those that preserve a monotonic hierarchical organisation during agglomeration, supporting the idea that protein configurational space has an approximately ultrametric, self-similar organisation. [Chapter 5](#), in turn, shifted the focus from the resolution of configurational states to the resolution of the structure itself, and showed that the optimal level of detail depends systematically on both system size and the quality of the conformational variability explored.

For learning networks, [Chapter 8](#) and [Chapter 9](#) studied two complementary aspects of neural networks: the organisation of their configurational space, and their structural properties in the course of training. [Chapter 8](#) focused on the network weight space and showed that the [Density of States](#) provides a global characterisation of the loss landscape, enabling analysis of how its structure depends on properties of the learning task, such as input statistics, linear separability, and class imbalance. [Chapter 9](#), in turn, shifted the focus from network weight configurations to hidden-layer representations; this chapter showed that [ME](#) identifies non-trivial subsets of hidden neurons that reflect the internal organisation of the learned representation and retain good predictive performance.

Overall, this thesis develops a unified perspective on complex systems, pivoting on the idea that useful descriptions strike a balance between reducing detail and preserving information. Across both proteins and learning networks, the objective is not simply to compress the fine-grained description, but to identify reduced representations that remain informative about the organisation of the system's configurational space. From this perspective, the main contribution of this work is twofold: on the one hand, it provides general methodological results for the study of complex systems; on the other, it supports a strong conceptual point, namely that low-resolution representations are more than approximations, but rather effective tools for revealing emergent structure and behaviour.

In my view, one of the most far-reaching implications of this work is that approaches originally developed within [CG](#) modelling may prove useful well beyond the scope for which they were first conceived. We live in an age of unprecedented

informational abundance, in which detailed descriptions of natural and artificial systems are becoming increasingly easier to generate but harder to process and understand. In such a context, the importance of low-resolution representations lies in their capacity to let information emerge. For this reason, I believe that the elegance and generality of **CG** approaches may prove essential not only for reducing complexity, but also, and perhaps more importantly, for making sense of it.

Appendices



Relevance-Resolution: Optimality Point Identification

Once the **RR** curve has been constructed, the next step is to identify the representative point corresponding to a prescribed local slope μ^* . In the applications discussed in this thesis, two values are of particular interest: $\mu^* = -1$, which defines the information-theoretic balance between gain in H_{rel} and cost in H_{res} , and $\mu^* = 0$, which corresponds to the maximum of H_{rel} . More generally, however, the procedure is independent of the specific choice of μ^* and can be applied to any target slope.

This identification is non-trivial because the **RR** relation is not available as a smooth analytical function. Rather, it is obtained as a finite set of points $\{(H_{\text{res},i}, H_{\text{rel},i})\}$, generated by a family of candidate representations indexed by a control parameter λ . Depending on the application, λ may denote, for example, the number of clusters in a partition, the number of retained variables in a reduced description, the granularity of a discretisation, or any other parameter controlling the resolution of the representation. In general, different realisations at the same value of λ may produce slightly different values of H_{res} and H_{rel} , so that the empirical curve appears as a cloud of points rather than as a regular line.

To extract a robust estimate of the slope, the cloud of points is first converted into a sequence of representative points through a binning and averaging procedure. In this work we consider two simple protocols, referred to as *uniform density* and *uniform spacing*. In both cases, the points are grouped according to their value of H_{res} , and each group is replaced by a single representative point whose relevance is given by the average of the H_{rel} values of the points falling in that bin. A schematic illustration of the two procedures is reported in [Figure A.1](#) and [Figure A.2](#).

In the *uniform density* protocol, the points are ordered by increasing H_{res} and divided into consecutive groups containing the same number D of points. The bin widths in H_{res} are therefore variable, but each representative point is obtained from

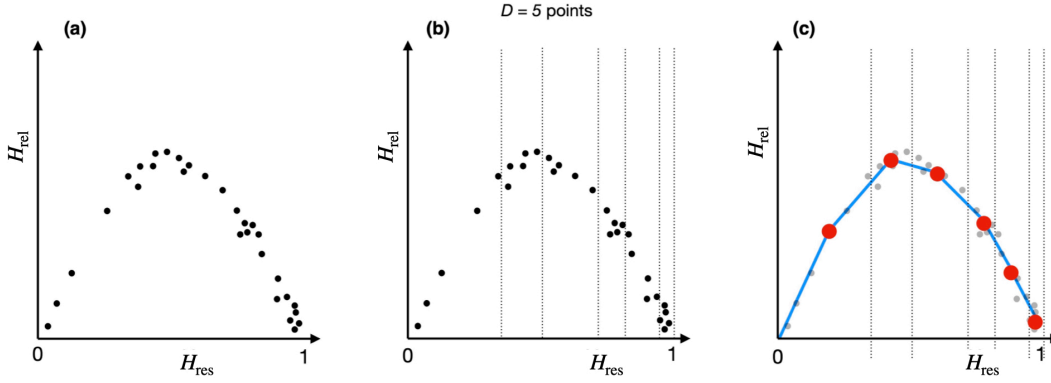


Figure A.1: Uniform density approach to binning resolution-relevance curves. (a) pictorial representation of 30 relevance and resolution points that we built ad hoc for a pictorial representation; (b) subdivision of the x axis into intervals of variable width, each one containing 5 points ($D = 5$); (c) average curve of relevance and resolution (red circles and blue lines).

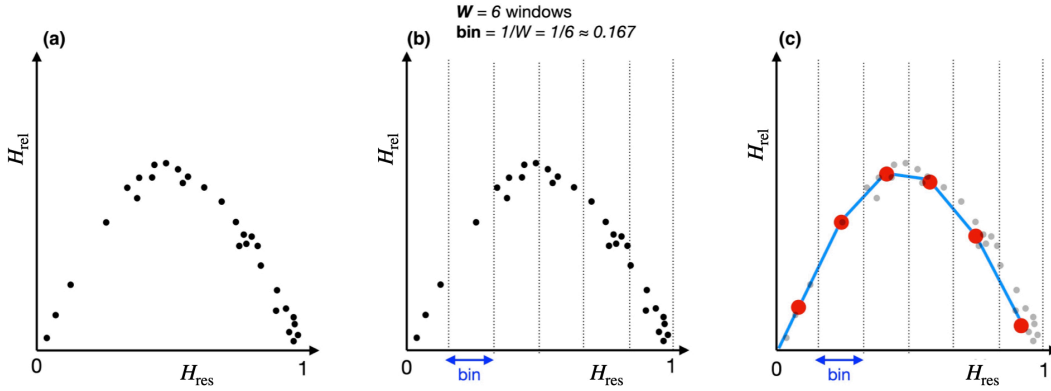


Figure A.2: Uniform spacing approach to binning resolution-relevance curves. (a) pictorial representation of 30 relevance and resolution points that we built ad hoc for a pictorial representation; (b) subdivision of the x axis into 6 intervals of same length ($W = 6$, bin width $1/W$), each one containing a variable number of points; (c) average curve of relevance and resolution (red circles and blue lines).

the same amount of data. This construction promotes a nearly uniform statistical weight along the averaged curve and is particularly useful when the original points are unevenly distributed in H_{res} . The procedure is illustrated in Figure A.1, where the points are grouped into bins containing a fixed number of elements and each bin is replaced by a representative averaged point.

In the *uniform spacing* protocol, instead, the H_{res} axis is subdivided into W intervals of equal width. The number of points in each bin is then variable, but the whole accessible range of H_{res} is covered uniformly by construction. This procedure is convenient when one wishes to preserve a homogeneous sampling of the resolution axis itself. A schematic example is shown in Figure A.2, where the averaging is performed over bins of fixed width along the H_{res} axis.

The outcome of either protocol is a sequence of representative points

$$\left(H_{\text{res}}^{(j)}, H_{\text{rel}}^{(j)}\right), \quad j = 1, \dots, n_{\text{bin}}, \quad (\text{A.1})$$

where $H_{\text{res}}^{(j)}$ is taken as the midpoint of the j -th bin and $H_{\text{rel}}^{(j)}$ is the arithmetic mean of the relevance values of the points belonging to that bin. These representative points, shown schematically as red markers in [Figure A.1](#) and [Figure A.2](#), define a smoothed version of the original **RR** scatter plot, suitable for numerical differentiation.

The local slope is then estimated from the representative points by finite differences. In the simplest implementation, used throughout this work, the derivative is computed through a first-order difference quotient,

$$\tilde{\mu}_j = \frac{H_{\text{rel}}^{(j+1)} - H_{\text{rel}}^{(j)}}{H_{\text{res}}^{(j+1)} - H_{\text{res}}^{(j)}}, \quad (\text{A.2})$$

with the slope $\tilde{\mu}_j$ associated with the interval between consecutive representative points. Equivalent higher-order finite-difference schemes could also be employed, but in practice the simple quotient was found to be sufficiently stable once the preliminary averaging step is applied.

Since the curve is finite and noisy, the condition $\tilde{\mu} = \mu^\star$ cannot be imposed exactly. Instead, one identifies all representative intervals whose slope falls within a tolerance window around the target value,

$$\tilde{\mu}_j \in [\mu^\star - \delta\mu, \mu^\star + \delta\mu]. \quad (\text{A.3})$$

In the calculations presented in this thesis, the default choice for the information-theoretic point is $\mu^\star = -1$ with tolerance $\delta\mu = 0.1$, namely $\tilde{\mu} \in [-1.1, -0.9]$. The same procedure is also used for the maximum-relevance point, corresponding to $\mu^\star = 0$, for which the accepted interval is $\tilde{\mu} \in [-0.1, 0.1]$. These intervals provide a reasonable compromise between numerical stability and fidelity to the target slope. An example of the full workflow for the case $\mu^\star = -1$ is shown in [Figure A.3](#).

Once the intervals satisfying the slope criterion have been identified, the associated values of the control parameter λ are collected. If several independent representations are available for the same λ , or if several neighbouring bins satisfy the condition, the procedure naturally yields a distribution of admissible values rather than a single one. The informative operating point is then defined from this distribution through its mean and standard deviation,

$$\lambda_{\text{opt}} = \langle \lambda \rangle, \quad \sigma_\lambda = \sqrt{\langle \lambda^2 \rangle - \langle \lambda \rangle^2}. \quad (\text{A.4})$$

In this way, the procedure returns both a central estimate of the optimal control parameter and a corresponding uncertainty that reflects the finite sampling of the

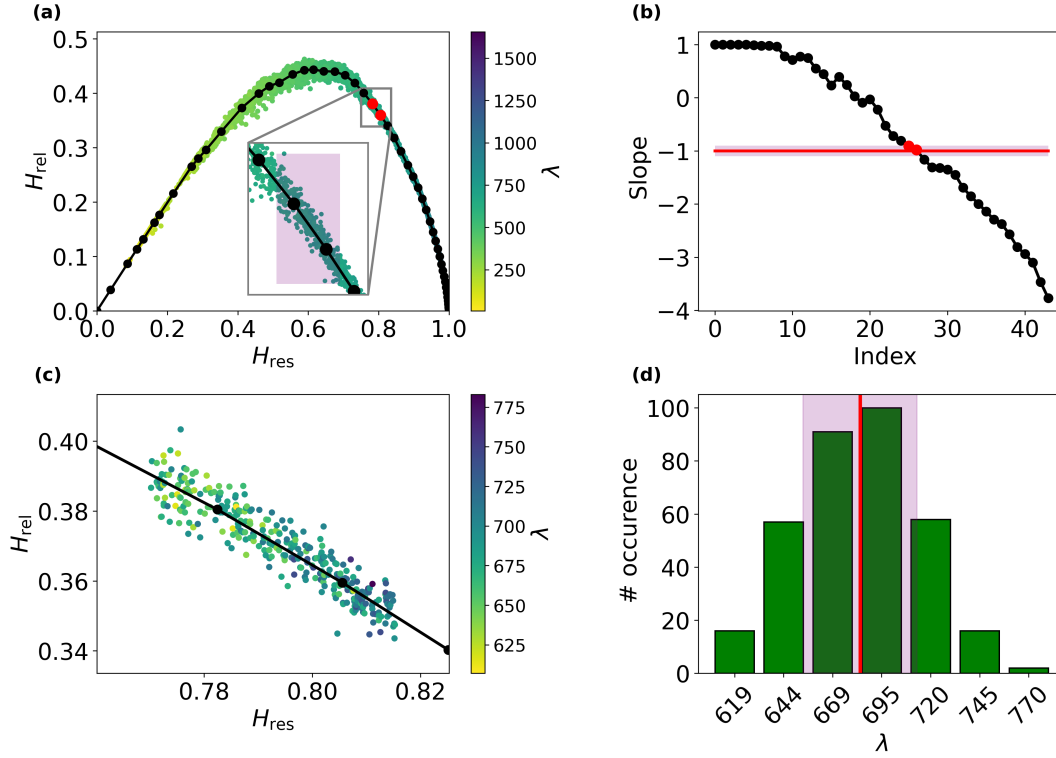


Figure A.3: Example of the workflow used to identify the informative operating point from an RR analysis, shown here for the target slope $\mu^* = -1$. **(a)** H_{rel} versus H_{res} scatter plot together with the smoothed curve obtained from the binning procedure. The highlighted region indicates the portion of the curve where the numerical slope falls within the prescribed tolerance window around the target value. **(b)** Enlargement of the selected region of the RR plot. **(c)** Numerical slope $\tilde{\mu}$ computed along the sequence of representative points; the horizontal band marks the tolerance interval $[-1.1, -0.9]$. **(d)** Distribution of the values of the control parameter λ associated with the intervals satisfying the slope criterion. The vertical line marks the mean value λ_{opt} , while the shaded region denotes one standard deviation around it. The same procedure is applied to the case $\mu^* = 0$, using the tolerance interval $[-0.1, 0.1]$.

RR curve and the residual variability of the candidate representations.

This construction is completely general and does not depend on the specific nature of λ . In the case of structural CG, $\lambda \equiv N_{\text{CG}}$, so that the above procedure yields the optimal number of retained sites and its uncertainty. In other contexts, λ may instead represent the number of clusters, the dimension of a reduced space, or any other parameter indexing the family of representations. The same algorithm can therefore be used whenever one wishes to identify, along an RR curve, the point corresponding to a prescribed information-theoretic condition.

As noted above, two operational definitions are used in this thesis. The first is the information-theoretic optimum, obtained from the condition $\mu^* = -1$; in the case of structural CG, the corresponding value is denoted by $N_{\text{OPT}}^{\text{IT}}$. The second is the maximum-relevance reference point, obtained from the condition $\mu^* = 0$; in the same context, the corresponding value is denoted by $N_{\text{OPT}}^{\text{MR}}$. The two estimates play complementary roles: $N_{\text{OPT}}^{\text{MR}}$ marks the coarsest representation at which H_{rel} is maximal, while $N_{\text{OPT}}^{\text{IT}}$ identifies the point where the marginal gain in H_{rel} balances

the marginal cost in H_{res} .

B

SI: Bliss of Dimensionality

This appendix gathers supplementary material associated with [Chapter 3](#). The results collected here extend the analyses presented in the main text and provide additional evidence for some of the points discussed there, while keeping the main exposition focused on the central methodological and conceptual aspects.

The appendix is organised as follows. [Section B.1](#) discusses the exclusion of mixtures involving digit 1 in the Gaussian-clone analysis of MNIST and illustrates the origin of the corresponding anomalous [RR](#) curves. [Section B.2](#) presents additional results for structured synthetic datasets, showing that the trends discussed in the main text remain robust across different parameter choices. Finally, [Section B.3](#) collects supplementary analyses on the behaviour of the [KL](#) criterion and clarifies, through explicit examples, the limitations of approaches based only on occupancy statistics.

B.1 Gaussian clones of MNIST: exclusion of digit 1

In the Gaussian-clone analysis of MNIST, digit 1 was excluded from the aggregate comparison because it induces a qualitatively different [RR](#) phenomenology. More precisely, for binary mixtures ($K = 2$) in which one of the two classes is digit 1, the [RR](#) curve is systematically not bell-shaped, but instead displays a robust bimodal structure with two distinct maxima of relevance. This behaviour falls outside the standard single-branch trade-off assumed in the interpretation of the [RR](#) curve and therefore makes the identification of the two characteristic optima ambiguous.

The effect is illustrated in [Figure B.1](#), where [RR](#) curves are reported for all digit pairs $(1, d)$, with $d \in \{0, 2, \dots, 9\}$. For each pair, the curve exhibits two peaks separated by a pronounced decrease in relevance at intermediate resolution. The important point is that this is not an isolated feature of a specific class combination, but a robust pattern shared by all tested pairs involving digit 1. In this regime, the theoretical motivation underlying the two operational optima introduced in the main text is

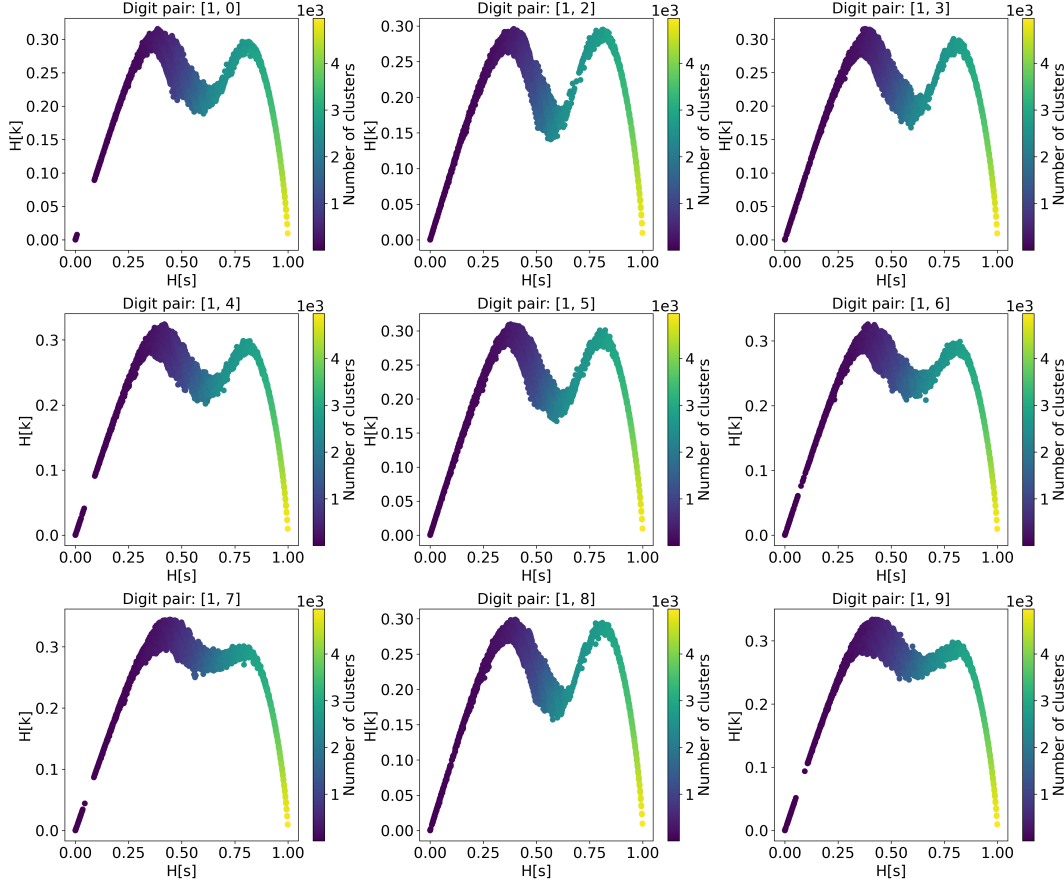


Figure B.1: RR curves for all digit pairs $(1, d)$, with $d \in \{0, 2, \dots, 9\}$, computed on MNIST-derived Gaussian-clone datasets. In each panel, H_{rel} is reported as a function of H_{res} for the family of representations generated by scanning the number of clusters. Points are coloured by the representation size parameter (as in the rest of the MNIST analysis).

no longer uniquely applicable. When the curve becomes multi-peaked, both multiple local maxima of relevance and multiple points satisfying the slope condition associated with the information-theoretic criterion may coexist on different branches of the curve. As a consequence, no intrinsic criterion within the framework allows one to privilege one branch over the other. In practice, both the maximum-relevance criterion and the slope-based criterion may select different branches depending on the specific realisation and on the local sampling density along the curve, leading to unstable estimates of $n_{\text{opt}}^{\text{MR}}$ and $n_{\text{opt}}^{\text{IT}}$ and to a deterioration of the comparison with n_{KL} .

A quantitative explanation for this anomalous behaviour is provided by the strong imbalance in the spatial extent of the class-conditional distributions in the pixel space. In pairs $(1, d)$, digit 1 occupies a much more concentrated region of feature space than the other digit. When the number of clusters is scanned, this introduces two distinct partitioning scales: a coarse regime in which the broader component is progressively resolved while the compact one remains effectively confined to a small number of states, and a finer regime in which the partition also begins to resolve the compact component. The coexistence of these two regimes is reflected

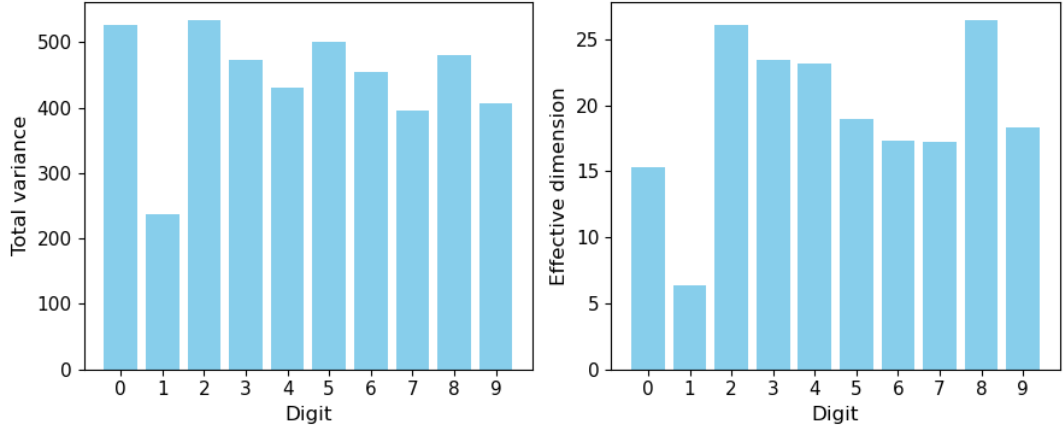


Figure B.2: Within-class scale and effective dimensionality across MNIST digits. Left: total within-class variance $V_c = \text{tr}(\hat{\Sigma}_c)$ computed from the class-conditional sample covariance $\hat{\Sigma}_c$ in $N = 784$ pixel space. Right: effective dimensionality $d_{\text{eff}}(c)$ defined by the participation ratio of the eigenvalues of $\hat{\Sigma}_c$.

by the appearance of two maxima of relevance along the **RR** curve.

To characterise this effect quantitatively, class compactness can be measured from the class-conditional covariance matrix $\hat{\Sigma}_c$ through the total within-class variance

$$V_c \equiv \text{tr}(\hat{\Sigma}_c) = \sum_{\alpha=1}^N \text{Var}(x_\alpha | c), \quad (\text{B.1})$$

and through the effective dimensionality, defined by the participation ratio of the eigenvalues $\{\lambda_i^{(c)}\}$ of $\hat{\Sigma}_c$,

$$d_{\text{eff}}(c) \equiv \frac{\left(\sum_i \lambda_i^{(c)}\right)^2}{\sum_i \left(\lambda_i^{(c)}\right)^2}. \quad (\text{B.2})$$

As shown in **Figure B.2**, digit 1 is an outlier under both descriptors, with substantially smaller within-class variance and lower effective dimensionality than the other digits. The imbalance in scale and effective dimensionality thus accounts for the coexistence of two distinct resolution regimes, and hence for the emergence of bimodal **RR** curves.

To test whether such a scale imbalance is sufficient, by itself, to generate the same phenomenology, one can consider a minimal *in silico* example based on a two-dimensional Gaussian mixture with components of different widths. In this case, the reference density is

$$p_{\text{ref}}(x) = \pi \mathcal{N}(x; \mu_1, \sigma_1^2 I_2) + (1 - \pi) \mathcal{N}(x; \mu_2, \sigma_2^2 I_2), \quad (\text{B.3})$$

with $x \in \mathbb{R}^2$ and $\sigma_1 \ll \sigma_2$. In the specific example shown in **Figure B.3**, the two components have same weights ($\pi = 0.5$) and different widths, with $\sigma_1 = 0.01$ and $\sigma_2 = 0.3$. In this setting, scanning the number of clusters again reveals two

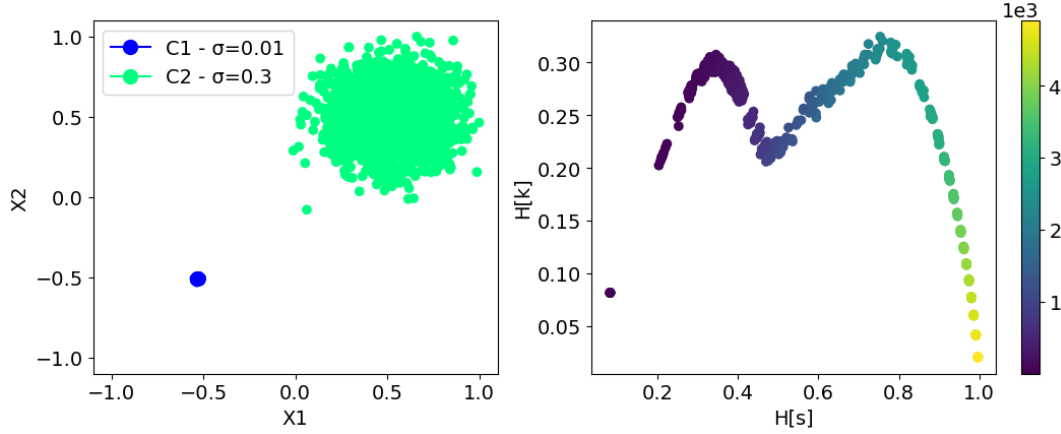


Figure B.3: In silico two-component mixture with scale imbalance and corresponding *RR* curve. Left: example realisation of samples $x \in \mathbb{R}^2$ drawn from a two-component Gaussian mixture with $\sigma_1 \ll \sigma_2$. Right: corresponding *RR* curve obtained by scanning the cluster size.

distinct resolution regimes: an initial one in which the broader component is progressively resolved while the narrower one remains effectively confined to a small number of states, and a second one in which the narrower component also begins to be resolved. The resulting rearrangement of occupancies produces two distinct maxima of relevance, thereby reproducing the qualitative mechanism observed in the MNIST mixtures involving digit 1.

Taken together, these results show that the exclusion of digit 1 is a necessary restriction to remain within the regime in which the usual *RR* interpretation is well defined. The standard construction of the two characteristic optima assumes a single regular trade-off between resolution and relevance. When strong scale and effective dimensionality imbalances generate multiple competing resolution regimes, this assumption breaks down, and the associated optimality criteria cease to be univocal.

B.2 Additional structured-synthetic parameter combinations

In this section, additional results for the structured synthetic datasets are reported in order to complement the analysis presented in the main text. The goal is to verify that the phenomenology discussed there is not tied to the specific parameter combinations shown in the corresponding figures, but remains stable across different choices of the number of mixture components, class weights, and relative widths of the informative and non-informative directions.

The results collected in Figure B.4 confirm the same qualitative trend observed in the main text. When the relevant structure is distributed over several informative directions, the optimal resolution identified by the *RR* analysis remains controlled by the amount of informative content carried by the data, rather than by the nominal

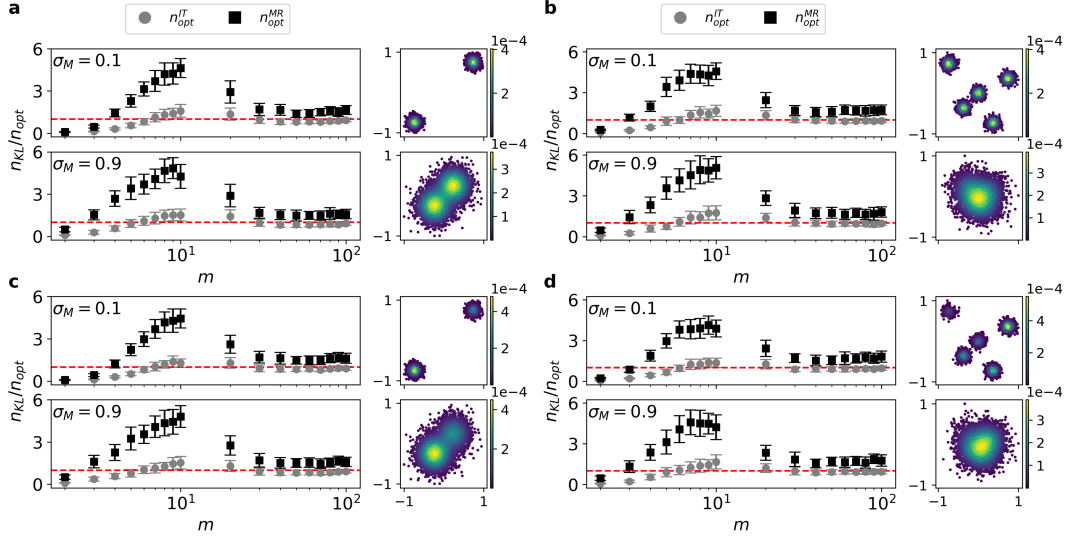


Figure B.4: Comparison between the KL-optimal representation and the two characteristic RR estimates for structured synthetic datasets. The figure shows the ratio $n_{\text{KL}}/n_{\text{opt}}$ as a function of the number of informative dimensions m , where n_{opt} denotes either the information-theoretic estimate $n_{\text{opt}}^{\text{IT}}$ (gray circles) or the maximum-relevance estimate $n_{\text{opt}}^{\text{MR}}$ (black squares). The dashed horizontal line marks the value 1, corresponding to perfect agreement with the KL optimum. Panels correspond to Gaussian mixture models with different numbers of components and mixing weights: (a) two components with equal weights, (b) five components with equal weights, (c) two components with unequal weights ($[0.66, 0.33]$), and (d) five components with unequal weights ($[0.34, 0.27, 0.19, 0.13, 0.07]$). For each case, results are shown for two values of the mixture width, $\sigma_M = 0.1$ and $\sigma_M = 0.9$. The plots on the right illustrate representative datasets in the corresponding informative subspaces.

dimensionality of the ambient space. Varying the number of classes, their relative statistical weights, or the widths associated with informative and noisy directions modifies the detailed shape of the curves, but does not alter the general picture.

These additional examples therefore support the interpretation proposed in the main text: the apparent blessing of dimensionality does not arise from dimensionality itself, but from the possibility of embedding the informative structure into a higher-dimensional representation without changing the effective number of degrees of freedom that are relevant for discrimination. In this sense, the RR criterion remains primarily sensitive to the organisation of the informative subspace, while the addition of non-informative dimensions has only a secondary effect.

B.3 Additional analyses on the KL criterion

The comparison discussed in the main text shows that the KL-based criterion provides a useful supervised benchmark for assessing the optimal resolution of a discrete representation. At the same time, the behaviour of the KL divergence also helps clarify an important limitation of criteria based solely on occupancy statistics. In this section, this point is made explicit through two complementary observations.

The first useful reference is provided by the two extreme representations. For

$n = 1$, all samples belong to a single state with occupancy $k_1 = P$, so that the induced empirical distribution over data points is uniform,

$$\hat{p}_{n=1}(x_i) = \frac{1}{P}. \quad (\text{B.4})$$

At the opposite extreme, for $n = P$, each sample forms a singleton state and all occupancies are equal to one. After point-wise normalisation, this again yields a uniform empirical distribution,

$$\hat{p}_{n=P}(x_i) = \frac{1}{P}. \quad (\text{B.5})$$

As a consequence, the **KL** divergence takes the same value at the two extremes,

$$D_{\text{KL}}(p \parallel \hat{p}_{n=1}) = D_{\text{KL}}(p \parallel \hat{p}_{n=P}) = \sum_{i=1}^P p(x_i) \log[P p(x_i)]. \quad (\text{B.6})$$

This identity is independent of the specific dataset and of the form of the reference distribution, and follows solely from the fact that both limiting representations induce a uniform empirical assignment over the observed samples.

For intermediate values of n , the partition generally produces a non-uniform distribution $\hat{p}(x_i)$, and the corresponding **KL** divergence may either decrease or increase relative to this uniform baseline. A decrease indicates that the representation assigns larger empirical weight to samples with larger reference probability, and therefore provides a better approximation of the target distribution than the uniform estimate. Conversely, an increase indicates that the partition is misaligned with the reference structure, so that the induced distribution is a worse approximation of $p(x_i)$ than the trivial uniform assignment.

This distinction can be illustrated explicitly in the structured synthetic setting, where samples $x = (u, v) \in \mathbb{R}^m \times \mathbb{R}^{N-m}$ are generated with latent structure in the informative subspace u and independent Gaussian noise in the remaining coordinates v . In this case, one can construct the clustering metric from three different sets of variables: the full feature vector x , only the informative coordinates u , or only the noisy coordinates v . The first case corresponds to the construction adopted in the main text, the second represents an idealised reference in which the informative variables are known in advance, and the third provides an explicit example of a representation built from variables that are uninformative with respect to the latent probability structure.

As shown in [Figure B.5](#), the behaviour of the **KL** divergence clearly discriminates between these three situations. When clustering is performed using the full feature set or the informative coordinates only, $D_{\text{KL}}(p \parallel \hat{p})$ decreases below the uniform baseline for intermediate resolutions, showing that the induced empirical distribution captures relevant aspects of the reference probability structure. By contrast, when clustering is performed using only the noisy coordinates, the **KL** divergence

increases above the baseline, indicating that the resulting state assignment is less consistent with the reference distribution than a uniform estimate.

Importantly, however, the corresponding **RR** curves remain regular and single-peaked in all three cases, including when the representation is built from noise-only features. This makes clear that the **RR** analysis depends only on the occupancy statistics induced by a given mapping from data to states, and not on whether that mapping resolves variables that are actually informative for the underlying system. As a consequence, a representation constructed from uninformative coordinates may still produce a standard bell-shaped **RR** curve, even though it yields a poor approximation of the target density according to the supervised **KL** benchmark.

The conclusion is therefore not that the **RR** framework fails, but rather that its scope must be interpreted correctly. The method can identify informative resolutions within a chosen family of representations, but it does not by itself validate the variables, distance measure, or clustering metric used to construct that family. These ingredients must already retain information that is relevant for the problem under study. Once this condition is met, the **RR** analysis provides a principled criterion to select the appropriate level of resolution.

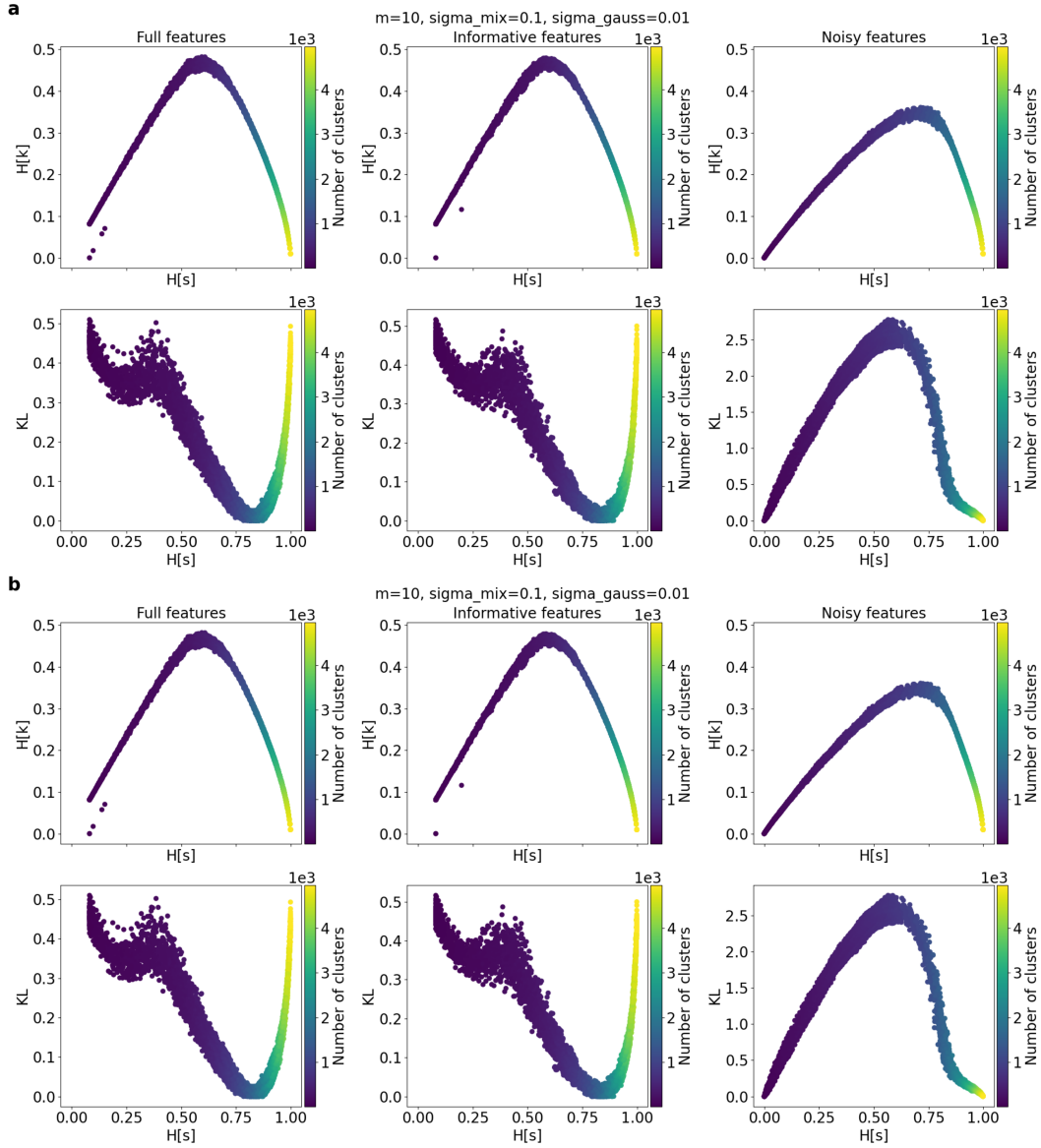


Figure B.5: Structured synthetic data ($K = 2$): RR and KL under different feature choices. Top row: RR curves obtained by clustering on the full feature set, on the informative coordinates only, and on the noisy coordinates only. Bottom row: corresponding KL divergence $D_{KL}(p||\hat{p})$ as a function of H_{res} for the same representations. Panels (a) and (b) correspond to $m = 5$ and $m = 10$, respectively, with $\sigma_M = 0.1$ and $\sigma_N = 0.01$.



SI: Configurational Coarse-Graining

This appendix collects supplementary material for the [Chapter 4](#). Its purpose is twofold: on the one hand, to provide additional analyses that support the interpretation of the adenylate-kinase results discussed in the main text; on the other hand, to report an additional validation on the humanised IgG4 monoclonal antibody system, which extends the comparison of clustering strategies beyond the case studies presented in the chapter.

The appendix is therefore organised into two parts. The first section presents additional [RR](#) analyses for adenylate kinase, focusing on quantities that help clarify the interpretation of the optimal clustering scale. The second section reports the results obtained for the humanised IgG4 monoclonal antibody and provides a complementary test of the robustness of the clustering procedures considered.

C.1 Additional analyses for adenylate kinase

This section presents additional analyses for the adenylate-kinase case study. In particular, it compares the representation that maximises the relevance with the one identified by the condition $\mu = -1$, and examines the behaviour of the inter-cluster distance as a function of the number of clusters.

The analysis was performed on the same adenylate-kinase trajectory discussed in the main text, using the coarse structural representation based on the C_β atoms and considering the same three classes of partitions: average linkage, single linkage, and random clustering. For each case, the same set of 22 low-resolution representations explored in the main text was considered. [Figure C.1](#) reports, together with the corresponding [RR](#) curves, the traces of the inter-state and intra-state covariance matrices and the Pearson correlation coefficient between the leading diffusion coordinates of the high-resolution and low-resolution representations. The repre-

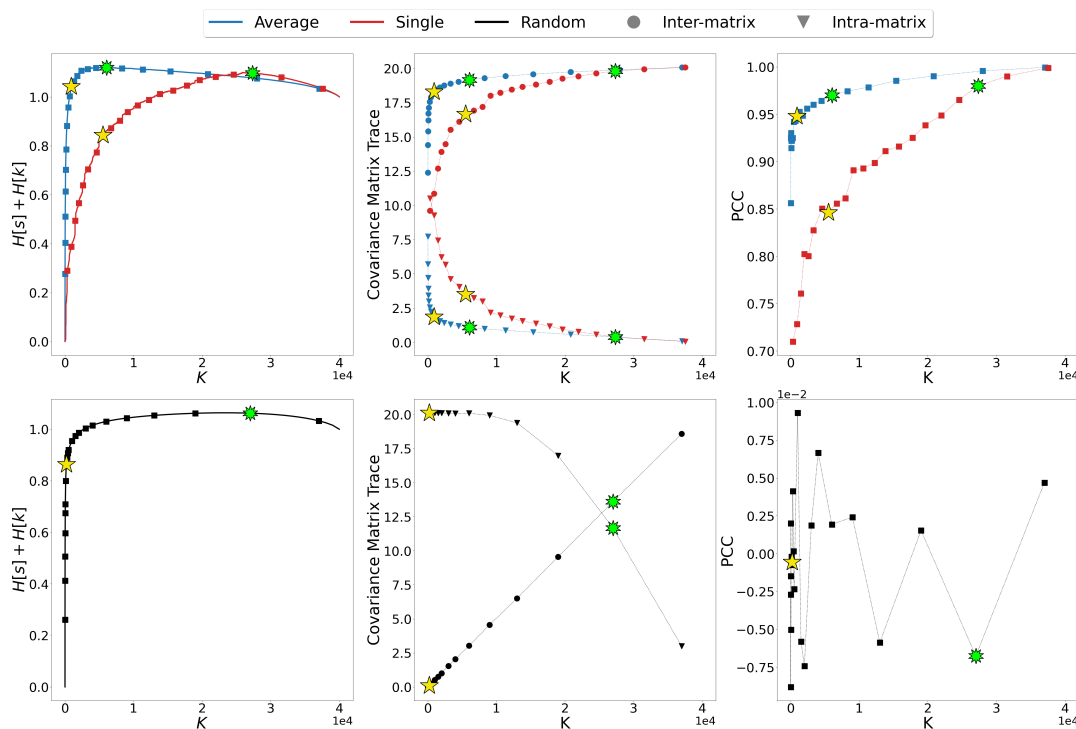


Figure C.1: Comparison between the characteristic representations identified along the RR curve for the adenylate-kinase trajectory. The left column shows the relevance-resolution curves, the central column the traces of the inter-state and intra-state covariance matrices, and the right column the Pearson correlation coefficient between the leading diffusion coordinates of the high-resolution and low-resolution representations, all as functions of the number of clusters K . The upper panels refer to average and single linkage clustering, while the lower panels report the corresponding random reference. Yellow stars mark the representations that maximise the relevance, whereas green stars indicate the representations that maximise $H[s] + H[k]$.

sensation that maximises the relevance is marked by a yellow star, while the one corresponding to $\mu = -1$ is marked by a green star.

For average linkage, the two characteristic points remain close to one another and both fall near the crossover region identified by the structural diagnostics. In the PCC, this corresponds to the onset of the rapid loss of agreement between the low-resolution and high-resolution diffusion spaces. In the covariance decomposition, it corresponds to the regime in which further reduction of the number of clusters begins to suppress an appreciable fraction of the inter-state component while increasing the intra-state one. This supports the interpretation that, beyond this scale, additional coarse graining mainly reduces structural information rather than only removing redundancy.

Single linkage displays a less clear behaviour. The separation between the relevance maximum and the point with $\mu = -1$ is broader, and neither of the two is associated with an equally sharp crossover in the PCC or covariance curves. This is consistent with the weaker performance of single linkage discussed in the main text, where its RR curve remains closer to the random reference and its low-resolution representations preserve the geometry of the high-resolution space less effectively

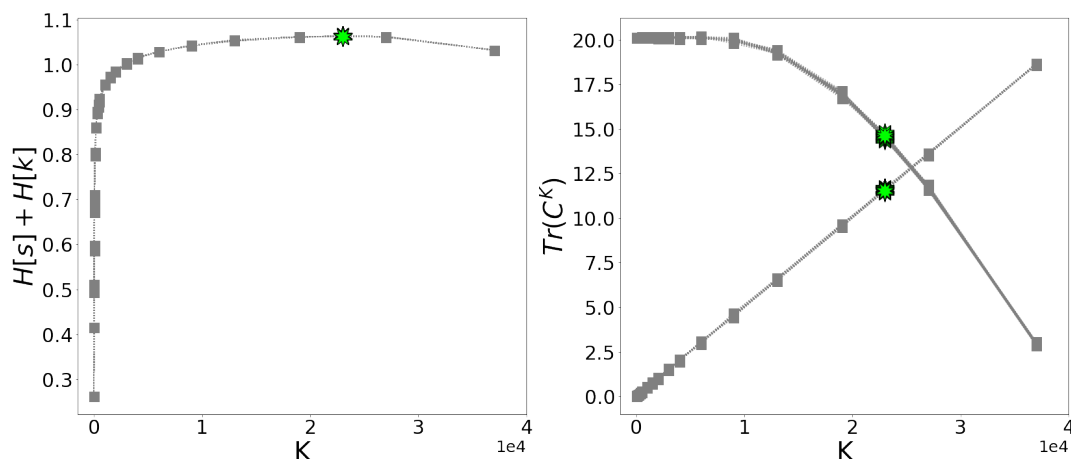


Figure C.2: Minimum inter-cluster distance as a function of the number of clusters K for the adenylate-kinase trajectory, for average linkage (left) and single linkage (right). Yellow stars mark the representations that maximise the relevance, while green stars identify those that maximise $H[s] + H[k]$.

than those obtained with average linkage.

The random reference helps clarify the meaning of these characteristic points. In this case, the point selected by $\mu = -1$ is not associated with an informative organisation of the configurational space, but rather with a regime in which the distinction between inter-state and intra-state variability is largely lost. The additional random realisations shown in Figure C.2 confirm that this behaviour is systematic.

A complementary perspective is provided by the behaviour of the inter-cluster distance as a function of the number of clusters, reported in Figure C.3. For both average and single linkage, the minimum distance required to obtain a partition into K clusters decreases rapidly at small K , and then more gradually as finer partitions begin to separate increasingly similar configurations. In both cases, the relevance maximum is located immediately after the main change of slope of this curve. The representation selected by the RR analysis therefore corresponds to the point at which the dominant separation scales in the configurational space have already been resolved, while further refinement mostly acts within regions that are already densely sampled.

Overall, these results show that, for adenylate kinase, the characteristic points identified from the RR analysis correspond to detectable crossovers in independent structural diagnostics, especially for average linkage. This further supports the interpretation of the selected low-resolution representation as a meaningful structural scale.

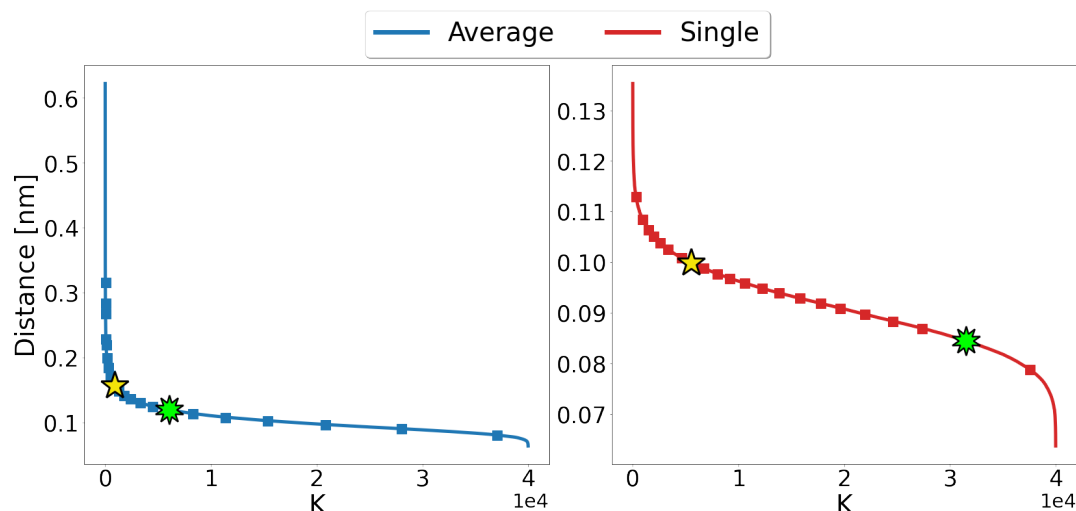


Figure C.3: Relevance-resolution curve and covariance-matrix traces for an additional random partition of the adenylate-kinase trajectory. The left panel reports the relevance-resolution curve as a function of the number of clusters K , while the right panel shows the traces of the inter-state and intra-state covariance matrices. The green stars identify those that maximise $H[s] + H[k]$.

C.2 Additional validation on the humanised IgG4 monoclonal antibody

As an additional validation of the analysis presented in the main text, we considered a humanised IgG4 monoclonal antibody, a system that is both larger and structurally more complex than adenylate kinase. The corresponding trajectory was generated over a timescale of 2 ms and analysed through the same configurational coarse-graining pipeline adopted for the other proteins. In this way, the IgG4 system provides a useful test of whether the conclusions drawn from the adenylate-kinase case study remain valid in a more complex conformational landscape.

Figure C.4 and Figure C.5 compare the low-resolution representations obtained for the antibody at the two values of K corresponding to the maxima of the relevance curves for complete and average linkage, respectively. As in the main text, the comparison is performed by relating each low-resolution representation to the geometry of the corresponding high-resolution reference. This allows one to assess whether the clustering procedures identified as most informative by the RR analysis also preserve the structural organisation of the configurational space in this larger system.

The results are consistent with the picture obtained for adenylate kinase. Average and complete linkage preserve the structure of the high-resolution space more faithfully than single linkage. In particular, their low-resolution centroids remain distributed in a way that is more consistent with the geometry of the reference trajectory, whereas single linkage yields a visibly poorer representation, with centroids arranged along much less structured patterns.

Overall, these results provide an additional validation of the conclusions of the

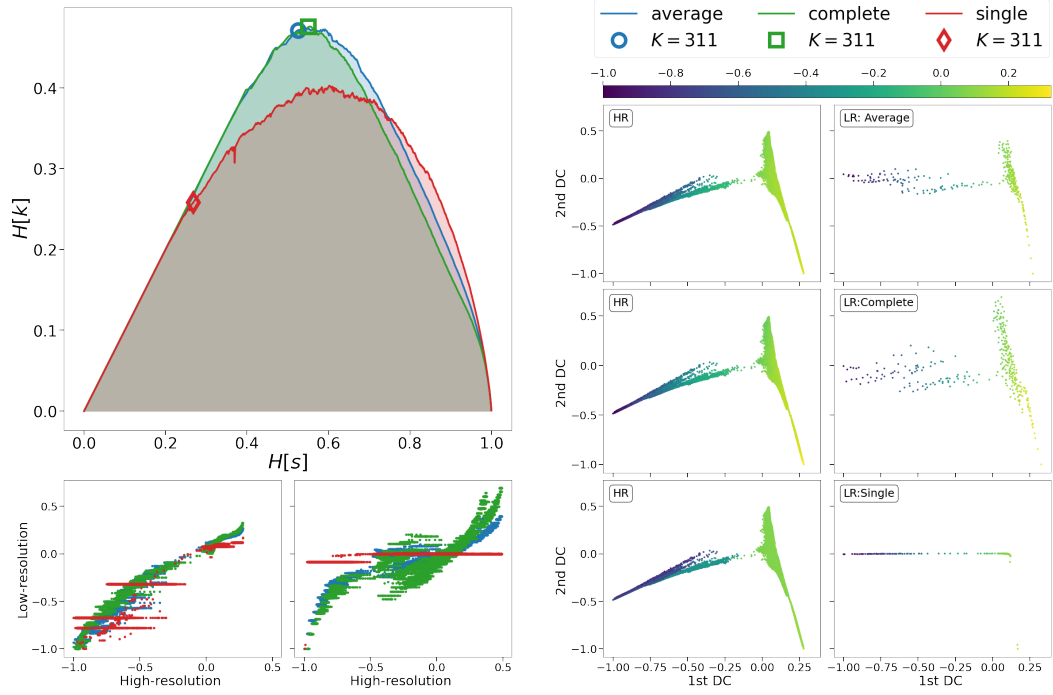


Figure C.4: Low-resolution representations of the humanised IgG4 monoclonal antibody at $K = 311$, corresponding to the maximum of the relevance curve for complete linkage. The left panel reports the relevance-resolution curves for average, complete, and single linkage, with the same value of K indicated on all three curves for comparison. The upper-right panels show the first and second diffusion coordinates of the high-resolution representation, while the lower-right panels show, for each clustering procedure, scatter plots in which the first and second diffusion coordinates of the high-resolution space are plotted against the corresponding coordinates of the low-resolution representation. Overall, average and complete linkage yield low-resolution representations that remain more consistent with the organisation of the high-resolution space than single linkage.

chapter. Also in the humanised IgG4 antibody, the clustering procedures that yield more informative low-resolution representations according to the RR analysis are those that better preserve the geometric organisation of the corresponding high-resolution space.

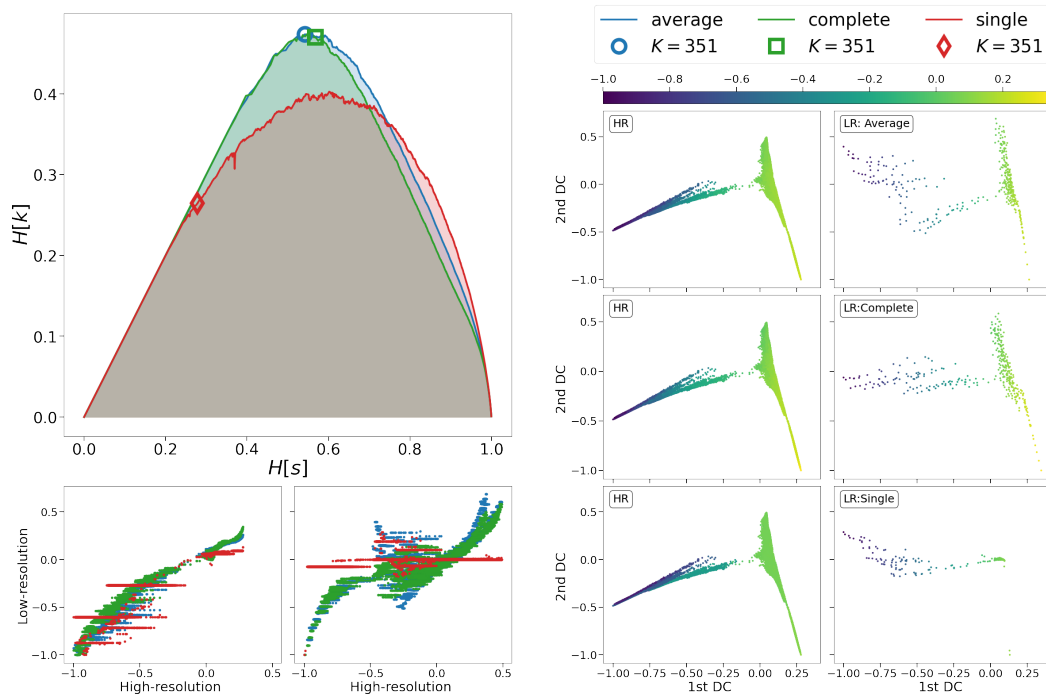


Figure C.5: Low-resolution representations of the humanised IgG4 monoclonal antibody at $K = 351$, corresponding to the maximum of the relevance curve for average linkage. The left panel reports the relevance-resolution curves for average, complete, and single linkage, with the same value of K indicated on all three curves for comparison. The upper-right panels show the first and second diffusion coordinates of the high-resolution representation, while the lower-right panels show, for each clustering procedure, scatter plots in which the first and second diffusion coordinates of the high-resolution space are plotted against the corresponding coordinates of the low-resolution representation. Overall, average and complete linkage yield low-resolution representations that remain more consistent with the organisation of the high-resolution space than single linkage.

D

SI: Structural Resolution

This appendix collects a set of supplementary analyses supporting the results presented in [Chapter 5](#). These materials are reported to complement the main discussion with additional quantitative comparisons and control analyses, while preserving the chapter’s continuity.

More specifically, the appendix is organised into three parts. The first part extends the analysis of the 11-protein dataset by considering, in addition to the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$ discussed in the main text, also the alternative estimate $N_{\text{OPT}}^{\text{MR}}$ obtained from the maximum of the relevance. The second part presents a more quantitative comparison of the structural resolutions identified by PROPRE with those associated with the Martini and SIRAH coarse-grained representations. The third part is devoted to robustness and reproducibility analyses, aimed at quantifying the sensitivity of the identified resolution to stochastic mapping generation and to the use of independent simulation replicas.

D.1 Extended fit on the 11-protein dataset

In the main text, the scaling analysis of the 11-protein dataset was discussed by focusing on the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$, corresponding to the point of slope -1 along the [RR](#) curve. This choice was motivated by the fact that $N_{\text{OPT}}^{\text{IT}}$ identifies the representation at which the marginal gain in relevance balances the marginal increase in resolution, and therefore provides the most natural operational definition of the optimal structural scale. At the same time, as discussed in [Appendix A](#), the same identification procedure can also be applied to the point $\mu = 0$, yielding the complementary estimate $N_{\text{OPT}}^{\text{MR}}$, which marks the coarsest representation at which the relevance is maximal.

[Figure D.1](#) extends the analysis presented in the chapter by reporting, for both the striding-based and clustering-based sampling strategies, the optimal number of retained sites as a function of the total number of residues for the full set of proteins,

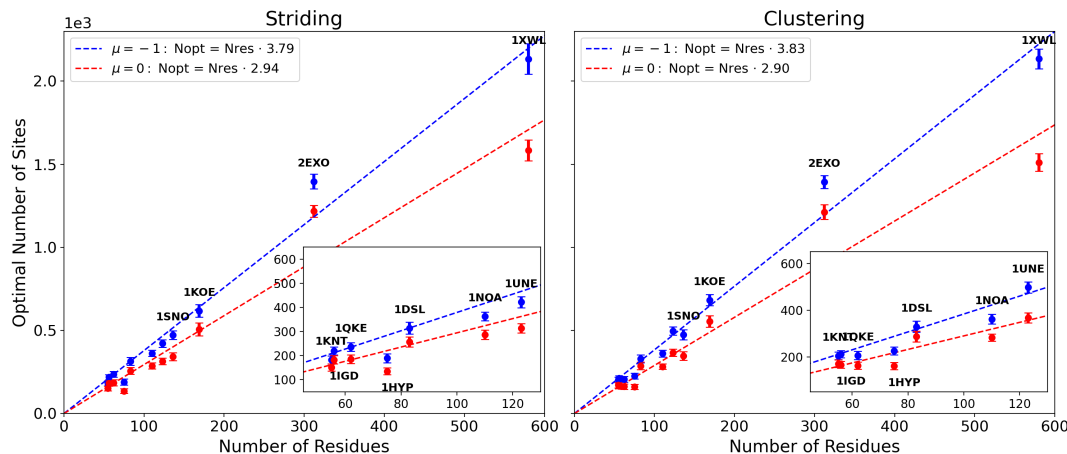


Figure D.1: Scaling of the optimal structural resolution with protein size for the 11-protein dataset. The optimal number of retained sites identified by PROPRE is reported as a function of the number of residues for the proteins considered in the dataset. The left panel refers to the striding-based sampling strategy, where 1000 equidistant frames are extracted from the trajectory, while the right panel refers to the clustering-based strategy, where 1000 cluster centroids are used. Blue points denote the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$, corresponding to the point of slope $\mu = -1$ along the RR curve, whereas red points denote the relevance-maximum estimate $N_{\text{OPT}}^{\text{MR}}$, corresponding to $\mu = 0$. Error bars indicate the uncertainty associated with the identification of the optimal interval. Dashed lines show the corresponding linear fits, with the reported proportionality constants between the optimal number of retained sites and the number of residues.

using both operational definitions. The figure shows that the approximately linear scaling between optimal resolution and protein size is preserved not only for $N_{\text{OPT}}^{\text{IT}}$, but also for $N_{\text{OPT}}^{\text{MR}}$.

The main difference between the two cases is quantitative rather than qualitative. While the fit corresponding to $N_{\text{OPT}}^{\text{IT}}$ yields a larger proportionality constant, the fit obtained from $N_{\text{OPT}}^{\text{MR}}$ is characterised by a smaller slope, reflecting the fact that the maximum of the relevance is reached at a coarser structural description. In practice, the fitted slope decreases from about 3.8 retained sites per residue at $\mu = -1$ to about 2.9 at $\mu = 0$, while maintaining similarly high correlation coefficients for both trajectory-sampling procedures.

The two estimates delimit a range of structurally meaningful resolutions for globular proteins: $N_{\text{OPT}}^{\text{IT}}$ identifies the finer end of this range, where structural information is preserved more faithfully, whereas $N_{\text{OPT}}^{\text{MR}}$ marks a coarser but still representative description, selected by the maximum of the relevance. This interpretation is consistent with the more general framework introduced in Section 2.3, where the two operating points are explicitly described as complementary rather than alternative definitions.

D.2 Quantitative comparison with Martini and SIRAH

The comparison with the Martini and SIRAH modelling, discussed in the main text, can be made more quantitative by considering, for each protein in the dataset, the

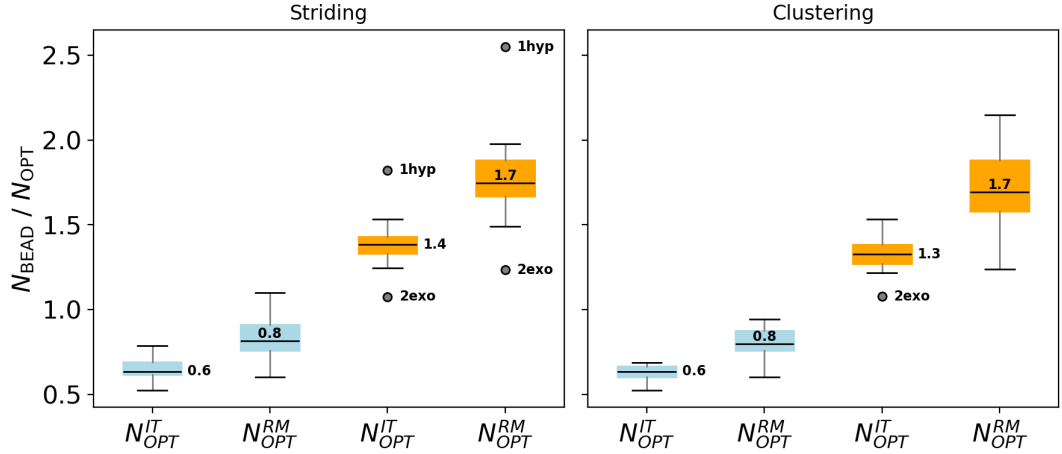


Figure D.2: Quantitative comparison between the structural resolutions identified by PROPRE and those associated with Martini-like and SIRAH-like representations. For each protein in the 11-protein dataset, the ratio between the number of coarse-grained beads used in Martini-like and SIRAH-like models and the optimal number of retained sites identified by PROPRE is reported. The left and right panels correspond to the striding-based and clustering-based sampling strategies, respectively. For each force field, two boxplots are shown: the first refers to the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$ and the second to the relevance-maximum estimate $N_{\text{OPT}}^{\text{MR}}$. Boxes indicate the interquartile range, the horizontal line marks the median, and individual outliers are labeled by protein pdb code. The numerical values reported above each box denote the median ratio.

ratio between the number of beads used in these two standard CG representations and the number of retained sites identified by PROPRE. Expressing the comparison in this form makes it possible to place different modelling strategies on a common relative scale, independently of the absolute protein size.

The analysis is performed with respect to both $N_{\text{OPT}}^{\text{IT}}$ and $N_{\text{OPT}}^{\text{MR}}$, and for both trajectory-sampling strategies considered throughout the chapter. As shown in Figure D.2, the trend is consistent across the entire protein set. In all cases, the Martini representation is associated with a smaller number of beads than the corresponding PROPRE estimate, whereas the SIRAH representation retains a larger number of sites.

Quantitatively, the Martini-to-PPRE ratio lies in the range 0.6-0.8, depending on the sampling method and on the operational definition of the PROPRE optimum. By contrast, the SIRAH-to-PPRE ratio falls in the range 1.3-1.7, indicating that SIRAH systematically adopts a finer representation. The structural resolution identified by PROPRE therefore lies between the two force-field schemes, supporting the interpretation advanced in the main text: the method does not reproduce one specific existing mapping, but selects an intermediate structural scale between the coarser Martini representation and the finer SIRAH representation.

This comparison should be interpreted at the level of resolution rather than of mapping. PROPRE does not aim to reconstruct the specific chemical grouping rules underlying transferable force fields, nor does it imply that information-theoretic considerations uniquely determine the mappings employed in Martini or SIRAH.

Rather, the result shows that the structural scale selected by PROPRE is consistent with the range of resolutions that has proved effective in practical CG modelling.

A further relevant aspect of Figure D.2 is that the relative ordering between the three resolutions is stable with respect to the precise operational definition of the PROPRE optimum. When the comparison is performed against $N_{\text{OPT}}^{\text{MR}}$, the ratios shift as expected because the reference resolution becomes coarser, but the overall conclusion remains unchanged: the Martini representation lies below the PROPRE range, whereas the SIRAH representation lies above it. This stability reinforces the interpretation of the PROPRE estimate as identifying in an unsupervised way a meaningful intermediate structural scale.

D.3 Robustness and Reproducibility of the PROPRE estimates

Since the PROPRE protocol relies on a stochastic exploration of the mapping space, it is necessary to assess the stability of the estimated informative structural resolution with respect to the finite sampling of candidate mappings. This point is particularly relevant because an exhaustive enumeration of all possible subsets of retained atoms is computationally unfeasible, owing to the combinatorial growth of the number of admissible mappings with the number of heavy atoms. For this reason, the practical implementation of the method necessarily relies on random sampling, and its robustness must be evaluated explicitly.

D.3.1 Robustness with respect to stochastic sampling of the mapping space

To quantify the sensitivity of the results to stochastic mapping generation, the full PROPRE analysis was repeated four times by varying the random realizations used to sample the retained-atom subsets at each CG level. The corresponding results are reported in Figure D.3 for three representative proteins.

The figure shows that the identified operating points remain stable across independent runs. Different realizations produce small variations in the detailed structure of the sampled representation space, but these fluctuations have only a minor effect on the estimated values of $N_{\text{OPT}}^{\text{IT}}$ and $N_{\text{OPT}}^{\text{MR}}$. In all cases, the between-run dispersion remains smaller than, or at most comparable to, the internal uncertainty associated with the identification of the optimal interval. This indicates that the finite stochastic sampling of mappings is sufficient to characterize the relevant region of the mapping space and does not introduce a significant bias in the estimate of the informative structural resolution.

This result shows that the PROPRE criterion is not controlled by isolated accidental mappings. Rather, the informative region emerges as a stable property of

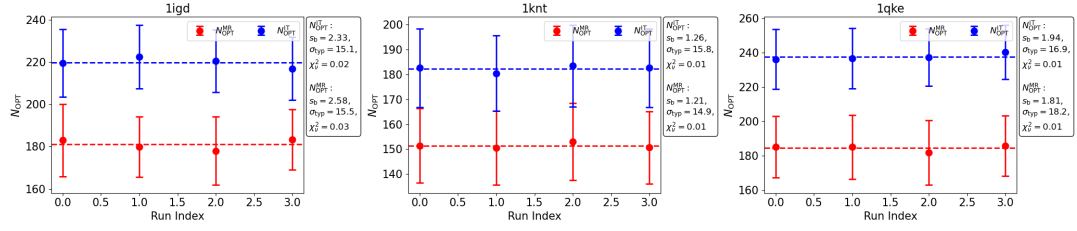


Figure D.3: Robustness of the PROPRE estimates with respect to stochastic sampling of the mapping space. Results obtained from four independent realizations of the PROPRE protocol for three representative proteins. For each realization, the estimated optimal number of retained sites N_{OPT} is reported with error bars corresponding to one standard deviation, while dashed lines indicate the mean across realizations. Blue markers denote the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$, corresponding to the point of slope $\mu = -1$, whereas red markers denote the relevance-maximum estimate $N_{\text{OPT}}^{\text{MR}}$, corresponding to $\mu = 0$. Insets report the between-run dispersion s_b , defined as the sample standard deviation of N_{OPT} across realizations, the typical internal uncertainty σ_{typ} , defined as the root-mean-square of the individual standard deviations, and the reduced chi-square χ^2_v testing consistency with a single underlying value.

the sampled ensemble of representations, and can therefore be identified reliably without requiring an exhaustive exploration of the full mapping space.

D.3.2 Reproducibility across independent simulation replicas

A second source of variability concerns the conformational ensemble used as input to the analysis. Even if the mapping-space exploration is stable, the estimated structural resolution could still depend on the particular molecular-dynamics trajectory from which the configurations are extracted. To assess this effect, the PROPRE analysis was repeated on independent simulation replicas. The corresponding results are shown in Figure D.4.

Also in this case, the identified structural resolution remains stable. Independent replicas lead to modest quantitative shifts in the location of the optimality interval, but they do not alter the overall position of the informative structural scale or its interpretation. The replica-to-replica variability is comparable to the intrinsic uncertainty of the method and therefore does not qualitatively affect the conclusions drawn in Chapter 5. This indicates that the estimated values of $N_{\text{OPT}}^{\text{IT}}$ and $N_{\text{OPT}}^{\text{MR}}$ reflect a genuine property of the conformational ensemble rather than an artifact of a specific trajectory realization.

Taken together, the analyses reported in Figure D.3 and Figure D.4 support the statistical reliability of the PROPRE protocol. The stochastic sampling of the mapping space is sufficiently dense to yield stable estimates of the relevant operating points, and the resulting structural resolutions are reproducible across independent realizations of the underlying molecular ensemble. The residual variability is naturally accounted for by the uncertainty associated with the operational identification of $N_{\text{OPT}}^{\text{IT}}$ and $N_{\text{OPT}}^{\text{MR}}$.

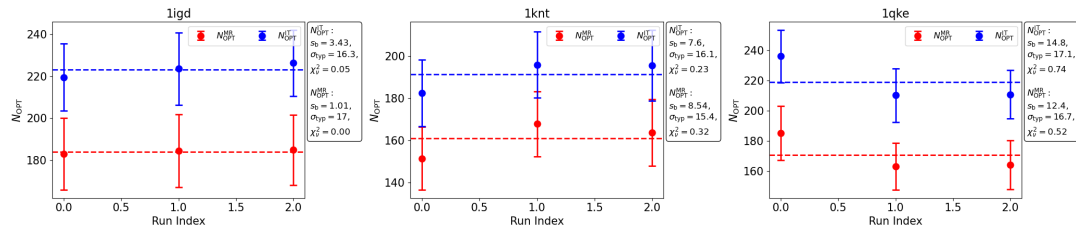


Figure D.4: Reproducibility of the PROPRES estimates across independent MD replicas. Results of PROPRES analyses performed on independent simulation replicas for three representative proteins. For each replica, the estimated optimal number of retained sites N_{OPT} is reported with error bars corresponding to one standard deviation, while dashed lines indicate the mean across replicas. Blue markers denote the information-theoretic estimate $N_{\text{OPT}}^{\text{IT}}$, corresponding to the point of slope $\mu = -1$, whereas red markers denote the relevance-maximum estimate $N_{\text{OPT}}^{\text{MR}}$, corresponding to $\mu = 0$. Insets report the between-replica dispersion s_b , defined as the sample standard deviation of N_{OPT} across replicas, the typical internal uncertainty σ_{typ} , defined as the root-mean-square of the individual standard deviations, and the reduced chi-square χ^2_v testing consistency with a single underlying value.

E

SI: Mapping Entropy Multi-Body Decomposition

This appendix reports the algebraic steps underlying the results presented in [Chapter 7](#). Its purpose is to make explicit the derivation leading to the multi-body expansions of the entropy. To this end, consider a system with n discrete degrees of freedom, collectively denoted by $\mathbf{x} \equiv (x_1, x_2, \dots, x_n)$, distributed according to the probability $p(\mathbf{x})$. Unless otherwise stated, all sums run over the full configurational space.

We introduced the characteristic function $\mathbb{I}(\mathbf{x}, \mathbf{x}')$, which takes the value 1 when the two configurations coincide and 0 otherwise. Using its factorised representation in terms of single-variable indicators,

$$\mathbb{I}(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^n \mathbb{I}(x_i, x'_i), \quad (\text{E.1})$$

and defining

$$\delta_i \equiv \mathbb{I}(x_i, x'_i) - 1, \quad (\text{E.2})$$

the probability associated with configuration $\mathbf{x}$ can be written as

$$p(\mathbf{x}) = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \mathbb{I}(\mathbf{x}, \mathbf{x}') = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \prod_{i=1}^n [1 + \delta_i] = \left\langle \prod_{i=1}^n [1 + \delta_i] \right\rangle, \quad (\text{E.3})$$

where $\langle \dots \rangle$ denotes the average over $\mathbf{x}'$ with weight $p(\mathbf{x}')$. Expanding the product gives

$$p(\mathbf{x}) = 1 + \sum_i \langle \delta_i \rangle + \sum_{i < j} \langle \delta_i \delta_j \rangle + \sum_{i < j < k} \langle \delta_i \delta_j \delta_k \rangle + \dots. \quad (\text{E.4})$$

Each term in this expansion remains a function of the reference configuration $\mathbf{x}$. In particular,

$$\langle \delta_i \rangle = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') [\mathbb{I}(x_i, x'_i) - 1], \quad (\text{E.5})$$

$$\langle \delta_i \delta_j \rangle = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') [\mathbb{I}(x_i, x'_i) - 1] [\mathbb{I}(x_j, x'_j) - 1], \quad (\text{E.6})$$

$$\langle \delta_i \delta_j \delta_k \rangle = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') [\mathbb{I}(x_i, x'_i) - 1] [\mathbb{I}(x_j, x'_j) - 1] [\mathbb{I}(x_k, x'_k) - 1]. \quad (\text{E.7})$$

Equation E.4 can be written in compact form as $p(\mathbf{x}) = 1 + X(\mathbf{x})$. The logarithm can then be expanded through the Taylor series. Taking terms up to the third order in X , one obtains

$$\begin{aligned} \ln p(\mathbf{x}) &= \ln(1 + X(\mathbf{x})) \\ &= \sum_i \langle \delta_i \rangle + \sum_{i < j} \langle \delta_i \delta_j \rangle + \sum_{i < j < k} \langle \delta_i \delta_j \delta_k \rangle \\ &\quad - \frac{1}{2} \sum_i \langle \delta_i \rangle^2 - \sum_{i < j} \langle \delta_i \rangle \langle \delta_j \rangle - \sum_{i < j} [\langle \delta_i \rangle \langle \delta_i \delta_j \rangle + \langle \delta_j \rangle \langle \delta_i \delta_j \rangle] \\ &\quad - \sum_{i < j < k} [\langle \delta_i \rangle \langle \delta_j \delta_k \rangle + \langle \delta_j \rangle \langle \delta_i \delta_k \rangle + \langle \delta_k \rangle \langle \delta_i \delta_j \rangle] \\ &\quad + \frac{1}{3} \sum_i \langle \delta_i \rangle^3 + \sum_{i < j} [\langle \delta_i \rangle^2 \langle \delta_j \rangle + \langle \delta_i \rangle \langle \delta_j \rangle^2] + 2 \sum_{i < j < k} \langle \delta_i \rangle \langle \delta_j \rangle \langle \delta_k \rangle + \dots \end{aligned} \quad (\text{E.8})$$

At this point is useful to rewrite the previous expansion grouping the different terms according to the number of distinct indices they involve

$$\begin{aligned} \ln(1 + X(\mathbf{x})) &= \sum_i \left[\langle \delta_i \rangle - \frac{1}{2} \langle \delta_i \rangle^2 + \frac{1}{3} \langle \delta_i \rangle^3 \right] \\ &\quad + \sum_{i < j} [\langle \delta_i \delta_j \rangle - \langle \delta_i \rangle \langle \delta_j \rangle - \langle \delta_i \rangle \langle \delta_i \delta_j \rangle - \langle \delta_j \rangle \langle \delta_i \delta_j \rangle + \langle \delta_i \rangle^2 \langle \delta_j \rangle + \langle \delta_i \rangle \langle \delta_j \rangle^2] \\ &\quad + \sum_{i < j < k} [\langle \delta_i \delta_j \delta_k \rangle - \langle \delta_i \rangle \langle \delta_j \delta_k \rangle - \langle \delta_j \rangle \langle \delta_i \delta_k \rangle - \langle \delta_k \rangle \langle \delta_i \delta_j \rangle + 2 \langle \delta_i \rangle \langle \delta_j \rangle \langle \delta_k \rangle] + \dots \end{aligned} \quad (\text{E.9})$$

Recalling the entropy definition:

$$S = - \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln p(\mathbf{x}) = - \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \ln(1 + X(\mathbf{x})), \quad (\text{E.10})$$

is possible to write it as

$$\begin{aligned}
S = & - \sum_i \left[\langle\langle \delta_i \rangle\rangle - \frac{1}{2} \langle\langle \delta_i^2 \rangle\rangle + \frac{1}{3} \langle\langle \delta_i^3 \rangle\rangle \right] \\
& - \sum_{i < j} \left[\langle\langle \delta_i \delta_j \rangle\rangle - \langle\langle \delta_i \rangle \langle \delta_j \rangle \rangle - \langle\langle \delta_i \rangle \langle \delta_i \delta_j \rangle \rangle - \langle\langle \delta_j \rangle \langle \delta_i \delta_j \rangle \rangle \right. \\
& \quad \left. + \langle\langle \delta_i^2 \delta_j \rangle\rangle + \langle\langle \delta_i \rangle \langle \delta_j^2 \rangle \rangle \right] \\
& - \sum_{i < j < k} \left[\langle\langle \delta_i \delta_j \delta_k \rangle\rangle - \langle\langle \delta_i \rangle \langle \delta_j \delta_k \rangle \rangle - \langle\langle \delta_j \rangle \langle \delta_i \delta_k \rangle \rangle - \langle\langle \delta_k \rangle \langle \delta_i \delta_j \rangle \rangle \right. \\
& \quad \left. + 2 \langle\langle \delta_i \rangle \langle \delta_j \rangle \langle \delta_k \rangle \rangle \right] + \cdots,
\end{aligned} \tag{E.11}$$

where $\langle\langle F(\mathbf{x}) \rangle\rangle \equiv \sum_{\{\mathbf{x}\}} p(\mathbf{x}) F(\mathbf{x})$ denotes the expectation value with respect to $p(\mathbf{x})$. To make the structure explicit, let us write some representative terms.

$$\begin{aligned}
\langle\langle \delta_i \rangle\rangle &= \sum_{\{\mathbf{x}\}} \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') p(\mathbf{x}) [\mathbb{I}(x_i, x'_i) - 1] \\
&= \sum_{x_i} \sum_{x'_i} p(x'_i) p(x_i) \mathbb{I}(x_i, x'_i) - 1 = \sum_{x_i} p(x_i)^2 - 1,
\end{aligned} \tag{E.12}$$

$$\begin{aligned}
\langle\langle \delta_i^2 \rangle\rangle &= \sum_{\{\mathbf{x}\}} p(\mathbf{x}) \left[\sum_{\{\mathbf{x}'\}} p(\mathbf{x}') (\mathbb{I}(x_i, x'_i) - 1) \right]^2 \\
&= \sum_{\{\mathbf{x}\}} \sum_{\{\mathbf{x}'\}} \sum_{\{\mathbf{x}''\}} p(\mathbf{x}) p(\mathbf{x}') p(\mathbf{x}'') [\mathbb{I}(x_i, x'_i) - 1] [\mathbb{I}(x_i, x''_i) - 1] \\
&= \sum_{x_i} \sum_{x'_i} \sum_{x''_i} p(x_i) p(x'_i) p(x''_i) \mathbb{I}(x_i, x'_i) \mathbb{I}(x_i, x''_i) - 2 \sum_{x_i} p(x_i)^2 + 1 \\
&= \sum_{x_i} p(x_i)^3 - 2 \sum_{x_i} p(x_i)^2 + 1,
\end{aligned} \tag{E.13}$$

$$\begin{aligned}
\langle\langle \delta_i \rangle \langle \delta_j \rangle \rangle &= \sum_{\{\mathbf{x}\}} \sum_{\{\mathbf{x}'\}} p(\mathbf{x}) p(\mathbf{x}') [\mathbb{I}(x_i, x'_i) - 1] [\mathbb{I}(x_j, x'_j) - 1] \\
&= \sum_{x_i, x_j} \sum_{x'_i, x'_j} p(x_i, x_j) p(x'_i, x'_j) \mathbb{I}(x_i, x'_i) \mathbb{I}(x_j, x'_j) - \sum_{x_i} p(x_i)^2 - \sum_{x_j} p(x_j)^2 + 1 \\
&= \sum_{x_i, x_j} p(x_i, x_j)^2 - \sum_{x_i} p(x_i)^2 - \sum_{x_j} p(x_j)^2 + 1,
\end{aligned} \tag{E.14}$$

$$\begin{aligned}
\langle\langle\delta_i\delta_j\rangle\rangle &= \sum_{\{\mathbf{x}\}} \sum_{\{\mathbf{x}'\}} p(\mathbf{x})p(\mathbf{x}') [\mathbb{I}(x_i, x'_i) - 1] [\mathbb{I}(x_j, x'_j) - 1] \\
&= \sum_{x_i, x_j} p(x_i, x_j)^2 - \sum_{x_i} p(x_i)^2 - \sum_{x_j} p(x_j)^2 + 1.
\end{aligned} \tag{E.15}$$

Using Equation E.11, S can be organised as a sum of contributions involving one, two, three, ... indices:

$$-S = \sum_i Q_i + \sum_{i < j} Q_{ij} + \sum_{i < j < k} Q_{ijk} + \dots \tag{E.16}$$

We now identify explicitly the first three terms.

One-body contribution. From Equation E.5 one has

$$1 + \langle\delta_i\rangle = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \mathbb{I}(x_i, x'_i) = p(x_i). \tag{E.17}$$

Therefore, all terms involving only the index i resum to

$$\langle\delta_i\rangle - \frac{1}{2}\langle\delta_i\rangle^2 + \frac{1}{3}\langle\delta_i\rangle^3 + \dots = \ln(1 + \langle\delta_i\rangle) = \ln p(x_i). \tag{E.18}$$

Averaging over $p(\mathbf{x})$ then yields

$$Q_i = \sum_{x_i} p(x_i) \ln p(x_i). \tag{E.19}$$

Two-body contribution. Using Equation E.5 and Equation E.6, one finds

$$1 + \langle\delta_i\rangle + \langle\delta_j\rangle + \langle\delta_i\delta_j\rangle = \sum_{\{\mathbf{x}'\}} p(\mathbf{x}') \mathbb{I}(x_i, x'_i) \mathbb{I}(x_j, x'_j) = p(x_i, x_j). \tag{E.20}$$

Therefore, the logarithm of the pair marginal can be written as

$$\ln p(x_i, x_j) = \ln(1 + \langle\delta_i\rangle + \langle\delta_j\rangle + \langle\delta_i\delta_j\rangle). \tag{E.21}$$

In order to isolate the genuine contribution associated with the pair (i, j) , one must subtract the one-body terms already assigned to i and j . Using Equation E.17, this gives

$$\ln p(x_i, x_j) - \ln p(x_i) - \ln p(x_j) = \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)}. \tag{E.22}$$

Expanding the previous expression up to third order in the averages, one obtains

$$\langle\delta_i\delta_j\rangle - \langle\delta_i\rangle\langle\delta_j\rangle - \langle\delta_i\rangle\langle\delta_i\delta_j\rangle - \langle\delta_j\rangle\langle\delta_i\delta_j\rangle + \langle\delta_i\rangle^2\langle\delta_j\rangle + \langle\delta_i\rangle\langle\delta_j\rangle^2 + \dots, \tag{E.23}$$

which coincides with the complete set of terms involving the pair (i, j) up to third

order. Averaging over $p(\mathbf{x})$, the corresponding two-body contribution is

$$Q_{ij} = \sum_{x_i, x_j} p(x_i, x_j) \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)}. \quad (\text{E.24})$$

Three-body contribution. Finally, consider the terms involving the triplet of indices (i, j, k) . Using Equation E.7, together with the lower-order identities above, one has

$$1 + \langle \delta_i \rangle + \langle \delta_j \rangle + \langle \delta_k \rangle + \langle \delta_i \delta_j \rangle + \langle \delta_i \delta_k \rangle + \langle \delta_j \delta_k \rangle + \langle \delta_i \delta_j \delta_k \rangle = p(x_i, x_j, x_k). \quad (\text{E.25})$$

Therefore, the contribution involving the triplet (i, j, k) is obtained by subtracting from $\ln p(x_i, x_j, x_k)$ all one-body and two-body pieces already accounted for. Explicitly,

$$\begin{aligned} & \ln p(x_i, x_j, x_k) - \ln p(x_i, x_j) - \ln p(x_i, x_k) - \ln p(x_j, x_k) \\ & \quad + \ln p(x_i) + \ln p(x_j) + \ln p(x_k) \\ &= \ln \frac{p(x_i, x_j, x_k) p(x_i) p(x_j) p(x_k)}{p(x_i, x_j) p(x_i, x_k) p(x_j, x_k)}. \end{aligned} \quad (\text{E.26})$$

Expanding the logarithms in powers of the averages, one recovers precisely the grouped three-body terms,

$$\langle \delta_i \delta_j \delta_k \rangle - \langle \delta_i \rangle \langle \delta_j \delta_k \rangle - \langle \delta_j \rangle \langle \delta_i \delta_k \rangle - \langle \delta_k \rangle \langle \delta_i \delta_j \rangle + 2 \langle \delta_i \rangle \langle \delta_j \rangle \langle \delta_k \rangle + \cdots, \quad (\text{E.27})$$

where again the omitted terms are the higher-order terms already contained in the full logarithmic resummation. Averaging over $p(\mathbf{x})$, the three-body contribution is therefore

$$Q_{ijk} = \sum_{x_i, x_j, x_k} p(x_i, x_j, x_k) \ln \frac{p(x_i, x_j, x_k) p(x_i) p(x_j) p(x_k)}{p(x_i, x_j) p(x_i, x_k) p(x_j, x_k)}. \quad (\text{E.28})$$

This identifies the first three terms of the entropy decomposition introduced in Chapter 7: the one-body contribution Q_i , the two-body contribution Q_{ij} , and the three-body contribution Q_{ijk} , given respectively by Equation E.19, Equation E.24, and Equation E.28.

E.1 Reference local measure and beyond the uniform degeneracy assumption

The one-body contribution admits a simple interpretation once the number of accessible states of each degree of freedom is written as

$$\omega_i = \frac{1}{p_i^0}, \quad (\text{E.29})$$

where p_i^0 denotes the probability for spin i to be found in any one of its accessible configurations in an unweighted and fully independent reference ensemble. Equivalently, p_i^0 is the uniform reference probability over the local state space of x_i .

Using Equation E.19, one then obtains

$$\begin{aligned} Q_i + \ln \omega_i &= \sum_{x_i} p(x_i) \ln p(x_i) + \ln \omega_i \\ &= \sum_{x_i} p(x_i) \ln p(x_i) + \sum_{x_i} p(x_i) \ln \omega_i \\ &= \sum_{x_i} p(x_i) \ln \frac{p(x_i)}{p_i^0}. \end{aligned} \quad (\text{E.30})$$

Since p_i^0 is uniform over the ω_i accessible states, the previous expression is precisely the Kullback–Leibler divergence between the empirical one-body marginal and the corresponding unweighted independent reference distribution,

$$Q_i + \ln \omega_i = D_{\text{KL}}(p(x_i) \parallel p_i^0). \quad (\text{E.31})$$

In particular,

$$Q_i + \ln \omega_i \geq 0, \quad (\text{E.32})$$

with equality if and only if $p(x_i) = p_i^0$ for all accessible states.

This interpretation also makes explicit the assumption underlying the ideal degeneracy term used in the main text. In the binary case, if the discarded degrees of freedom are assumed to be independent and each local state is taken to be equally likely in the reference ensemble, then $p_i^0 = 1/2$ and therefore $\omega_i = 2$ for every discarded spin. If $n - N$ variables are discarded, the corresponding degeneracy factor is

$$\Omega = 2^{n-N}, \quad (\text{E.33})$$

which gives the standard contribution $(n - N) \ln 2$.

A simple way to go beyond this assumption is to retain the independence of the discarded variables while relaxing the requirement that the two local states be equally weighted. Consider therefore a binary non-interacting degree of freedom x_i , with probability p_i of being in the up state and probability $1 - p_i$ of being in the

down state in the reference ensemble. The contribution of this variable is then no longer $\ln 2$, but rather the entropy of the corresponding Bernoulli distribution,

$$h(p_i) = -p_i \ln p_i - (1 - p_i) \ln(1 - p_i). \quad (\text{E.34})$$

Indeed, $h(p_i) = \ln 2$ only in the unbiased case $p_i = 1/2$, while $h(p_i) \rightarrow 0$ when $p_i \rightarrow 0$ or $p_i \rightarrow 1$, namely when the variable becomes effectively deterministic.

More generally, for a set D of discarded binary variables, assumed independent but not necessarily unbiased, the reference probability of a discarded configuration $\mathbf{y} = (y_i)_{i \in D}$ is

$$P_0(\mathbf{y}) = \prod_{i \in D} p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (\text{E.35})$$

with $y_i \in \{0, 1\}$. The corresponding reference entropy is

$$\begin{aligned} H_0(D) &= - \sum_{\mathbf{y}} P_0(\mathbf{y}) \ln P_0(\mathbf{y}) \\ &= \sum_{i \in D} [-p_i \ln p_i - (1 - p_i) \ln(1 - p_i)] \\ &= \sum_{i \in D} h(p_i). \end{aligned} \quad (\text{E.36})$$

Accordingly, the ideal degeneracy contribution

$$(n - N) \ln 2, \quad (\text{E.37})$$

is replaced by

$$\sum_{i \in D} h(p_i). \quad (\text{E.38})$$

This provides a simple extension of the ideal framework: instead of pure counting of uniformly accessible discarded states, one uses the entropy of the independent reference measure assigned to the discarded sector. The formal multi-body decomposition derived above is unchanged, but the balance between entropy and degeneracy in S_{map} is modified. In particular, when some discarded variables are strongly biased, their contribution is smaller than $\ln 2$, so that the effective degeneracy term is reduced with respect to the uniform prediction. This offers a natural route to rationalise deviations from the ideal behaviour when the assumption of equally weighted discarded states is no longer appropriate.

F

SI: Neural Network Configuration Space

This appendix collects supplementary material for [Chapter 8](#), including additional results on MNIST and Fashion-MNIST classification tasks, and a detailed description of the synthetic and Gaussian-clone datasets employed in the main text.

This appendix collects supplementary material for [Chapter 8](#). It includes additional classification results on MNIST and Fashion-MNIST tasks, presents the detailed construction of the synthetic datasets and Gaussian-clone benchmarks used in the analysis, and reports further technical aspects of the [WL](#) sampling procedure, including implementation details, validation tests, and scaling behaviour.

F.1 Additional details on the datasets

F.1.1 Synthetic data

This section provides a comprehensive description of the *synthetic* datasets used in this study, designed to model isotropic distributions in high-dimensional space. These datasets are structured to represent two distinct classes with configurable angular separation and inter-class distances, allowing controlled manipulation of their mean vector orientations and spatial relationships. The flexibility in specifying the parameters λ and $\Delta\mu$ enables various configurations, from parallel to orthogonal or antiparallel class centres. This setup allows for precise examination of class separability and interaction under different configurations, which is essential for assessing algorithmic performance in scenarios with high-dimensional, overlapping distributions.

The *synthetic* datasets represent isotropic distributions in N dimensions, with two classes whose mean vectors are derived from orthogonal normalized vectors sampled from a univariate normal distribution. Specifically, consider two such vectors in N dimensions, with components sampled from a univariate normal distri-

bution:

$$m_c = \left\{ \frac{m_i^c}{\|m_c\|} \right\}_{i=1}^N, \quad \text{with } m_i^c = \mathcal{N}(0, 1) \text{ and } c \in \{1, 2\}. \quad (\text{F.1})$$

The mean vectors of the two classes are defined as follows:

$$\begin{cases} \mu_1 \equiv \Delta\mu \cdot m_1 & \text{class 1} \\ \mu_2 \equiv \Delta\mu \cdot [\lambda \cdot m_1 + (1 - |\lambda|) \cdot m_2] & \text{class 2} \end{cases} \quad (\text{F.2})$$

Here, $\lambda \in [-1, 1]$ and $\Delta\mu \in [0, +\infty)$ are the morphing parameters controlling the angle between the two mean vectors and the inter-class separation, respectively. When $\lambda = 1$, the two vectors are parallel ($\mu_1 = \mu_2$) and the inter-class separation is zero ($\|\mu_1 - \mu_2\| = 0$). For $\lambda = 0$, the two vectors are orthogonal since $m_1 \perp m_2$, and the inter-class distance is $\|\mu_1 - \mu_2\| = \Delta\mu\sqrt{2}$. Finally, when $\lambda = -1$, the two mean vectors are antiparallel ($\mu_1 = -\mu_2$) and the inter-class separation is $\|\mu_1 - \mu_2\| = 2\Delta\mu$. For a generic value of λ , the inter-class distance is given by:

$$\|\mu_1 - \mu_2\| = \Delta\mu\sqrt{(1 - \lambda)^2 + (1 - |\lambda|)^2}. \quad (\text{F.3})$$

The angle between the mean vectors is defined by:

$$\theta = \arccos \frac{\mu_1 \cdot \mu_2}{\|\mu_1\| \|\mu_2\|} = \arccos \frac{\lambda}{\sqrt{\lambda^2 + (1 - |\lambda|)^2}}. \quad (\text{F.4})$$

F.1.2 Gaussian Clones

This section aims to report the methodology used to generate *Gaussian clones*, a family of synthetic datasets designed to approximate real-world data while allowing precise control over specific statistical properties. These datasets enable systematic exploration of how features such as mean separation and covariance structure influencing DoS properties. Two distinct types of Gaussian clones were constructed: the *Gaussian Mixture (GM) clone*, which retains both the mean vector (μ_c) and the full covariance matrix (Σ) of each class, and the *Isotropic Gaussian Mixture (2isoGM) clone*, which simplifies the covariance structure by preserving only the mean vectors and setting the variance (v) equal across classes.

For both variants, data were sampled from a multivariate Gaussian distribution $\mathcal{N}(\mu_c \Delta\mu, \Sigma)$, where μ_c is the class-specific mean vector in an N -dimensional space, $\Delta\mu$ is a scalar parameter controlling inter-class separation, and Σ is the covariance matrix. The covariance matrix was defined as:

$$\Sigma_{ij} = \begin{cases} \langle x_i - \langle x_i \rangle \rangle \langle x_j - \langle x_j \rangle \rangle & \text{for GM clones,} \\ v \delta_{ij} & \text{for 2isoGM clones.} \end{cases}$$

Here, δ_{ij} is the Kronecker delta, and $v = \sqrt{v_1 v_2}$ is the geometric mean of the variance of the two classes. The GM clone captures the real feature correlations of the data by preserving the full covariance matrix, while the 2isoGM clone eliminates these correlations, focusing instead on isotropic variance and the effects of mean separation. This approach provides a controlled framework for studying how statistical properties of the data shape algorithmic performance.

F.2 Additional details on the Wang-Landau algorithm

This section provides an in-depth overview of the Wang-Landau (WL) sampling method implementation, focusing on the parameters and conditions essential for its convergence and efficiency. The WL algorithm is an effective approach for high-dimensional sampling, dynamically adjusting probabilities to enable uniform visitation of energy levels and accurate reconstruction of a system's DoS. We detail the selection of local and global moves, chosen probabilistically to ensure thorough exploration across energy states, as well as the histogram flatness criterion, which is critical for reliable convergence and controls the reduction of the modification factor F . These implementation details are vital for replicating our simulations and understanding the precision of the WL method in our study. A pseudocode summary of the algorithm is included to outline the main steps and parameter choices.

At each stage of the WL simulation, the nature of the move—local or global—is determined by a random number $c \in (0, 1)$. Specifically, if $c < 0.8$, the move is local, meaning that one of the N components of the weight vector W is randomly selected and flipped to generate W' . Conversely, when $c > 0.8$, another random number is drawn to determine the number $n \leq N$ of components to be changed. These n components are then randomly selected and flipped.

The flatness condition is a crucial aspect of the WL algorithm, ensuring the uniformity of the histogram of visited energy levels. This condition is typically defined by the requirement that for each value of energy E , the histogram value $H_k(E)$ must not deviate significantly from the mean value $\langle H_k \rangle$. Mathematically, this is expressed as:

$$p_{flat} \times \langle H_k \rangle < H_k(E) < (2 - p_{flat}) \times \langle H_k \rangle,$$

where p_{flat} is a predefined flatness parameter.

In the simulations performed in our study, we set p_{flat} to 0.9. The flatness condition for the visited energy levels was checked every 800 Monte Carlo (MC) steps. The iterative simulation scheme was continued until the modification factor F decreased below the predefined final value, $F_{end} = \ln(f_{end}) = 10^{-6}$.

To illustrate the implementation of the Wang-Landau algorithm in our simulations, we provide the following pseudo-code outlining the key steps of the iterative process.

```

1 Initialize:
2
3     S(E) = 0 for all E //Initial entropy estimate for all energy levels
4     H(E) = 0 for all E //Initial histogram for all energy levels
5     V(E) == 0 for all E //Initial visited flag for all energy levels
6     Set modification factor F = 1
7     Initialize network state W
8     Compute initial energy E_current
9
10
11 while F > F_end:
12     //Perform n_MC Monte Carlo steps
13     for i = 1 to n_MC:
14
15         //Determine move type (local or global)
16         Generate random number  $c \in (0,1)$ 
17
18         if c < 0.8:
19             //Local move: Flip one random component
20             Randomly select 1 component of W
21             Flip it to generate W_trial
22
23         else:
24             //Global move: Flip  $n < N+1$  random components
25             Generate random number  $n \in [1,N]$ 
26             Randomly select n components of W
27             Flip them to generate W_trial
28
29         Compute trial energy E_trial
30
31         if E_trial within energy range:
32             //Check if this is a new energy level (first time visited)
33             if V(E_trial) == 0:
34                 Reset H(E) = 0 for all E //Reset histogram
35                 V(E_trial) = 1
36                 break //Restart the MC moves (break the loop)
37             end if
38
39         if S(E_current) > S(E_trial):
40             Accept with probability  $\exp(-(S(E\_trial) - S(E\_current)))$ 
41

```



```

42         else:
43             Accept the move
44
45             Update E_current
46             Update S(E_current) <- S(E_current) + F
47             Update H(E_current) <- H(E_current) + 1
48
49         end if
50     end for
51
52     //Check histogram flatness condition
53     Compute mean histogram value H_mean = average(H(E) for all E)
54
55     flatness_condition_met = True
56
57     for each energy level E:
58         if H(E) < p_flat * H_mean or H(E) > (2 - p_flat) * H_mean:
59             flatness_condition_met = False
60             break //Exit loop if flatness condition is not met
61
62     if flatness_condition_met:
63         Reset H(E) = 0 for all E //Reset histogram
64         Set F = F / 2 //Halve the modification factor
65
66     end if
67 end while
68
69 Output:
70     S(E), the estimated entropy for each energy level E

```

F.3 Validation and scaling analysis

Through a rigorous application of the Wang-Landau (WL) sampling algorithm with well-defined parameters, our approach ensures a thorough exploration of the network’s energy landscape, leading to precise estimations of the DoS , $\Omega(E)$, for the studied neural network.

In this section, we assess the accuracy of entropy estimations obtained via WL sampling for small neural networks. Additionally, we conduct a scaling analysis to examine how the convergence time of the algorithm grows with increasing network size and task complexity.

F.3.1 Validation of entropy estimation

To validate the accuracy of entropy estimations produced by WL sampling, we analyse small neural network architectures where an exhaustive search over the entire configuration space is feasible. This allows us to compute exact microcanonical entropy curves $S_T(E)$ and compare them with those estimated by WL sampling, denoted $S_{WL}(E)$.

Three network architectures were tested: (i) a single-layer perceptron with 15 input neurons, (ii) a one-hidden layer network with 5 input neurons and 3 hidden neurons for binary classification (one output neuron), and (iii) a one-hidden layer network for multi-class classification with 3 input neurons, 3 hidden neurons and 3 output neurons.

For each architecture, a dataset of P examples was generated using an identical teacher network: $P = 15$ for the single-layer perceptron, $P = 18$ for the binary classification network, and $P = 15$ for the multi-class classification network. To evaluate consistency, we performed 500 independent WL simulations per configuration, and the resulting entropy curves are shown in [Figure F.1](#).

The figure’s top panel displays the exact entropy curves (continuous red line) alongside the 500 WL-sampled curves (black points), indicating strong agreement. We quantified this agreement by computing the percentage difference between the sampled and exact curves using:

$$\text{dist}(S_{WL}(E), S_T(E)) = \frac{|S_{WL}(E) - S_T(E)|}{S_T(E)}. \quad (\text{F.5})$$

The mean and standard deviation of these distances, presented in the bottom panel of [Figure F.1](#), show that the average discrepancy remains consistently below $7e-2$. This outcome reinforces the reliability of WL sampling in estimating entropy curves with high accuracy.

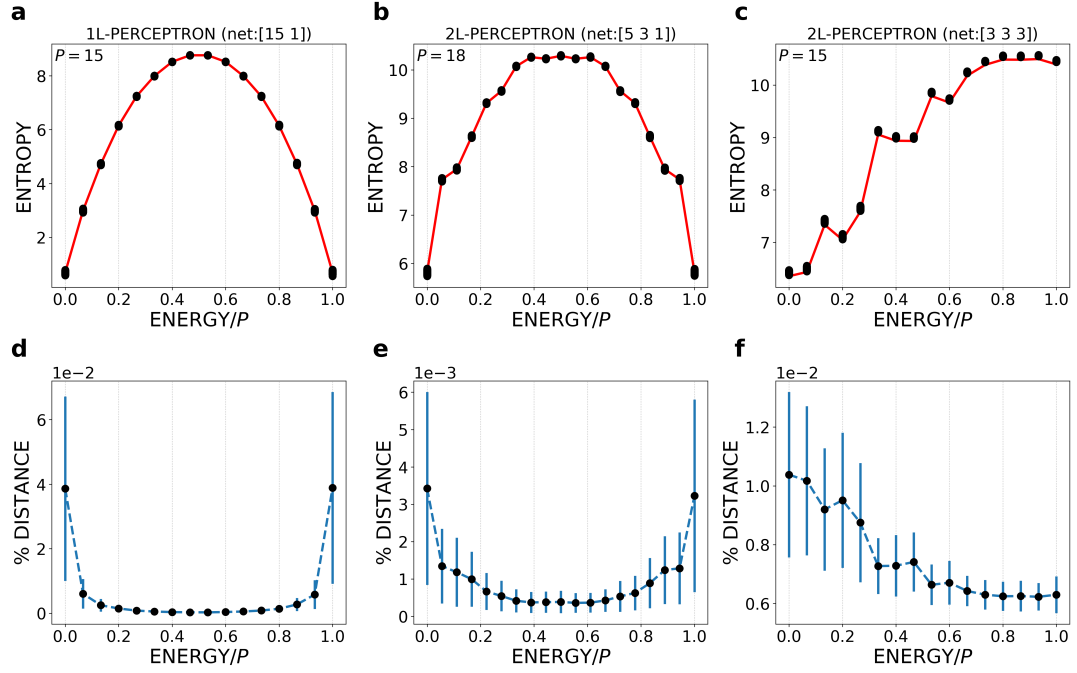


Figure F.1: Comparison of the true entropy curve to the output of the Wang-Landau algorithm across different network architectures. For each architecture, 500 independent simulations are performed on the same dataset composed of P data points. The upper panels (a-c) compare the true entropy curve (red-solid line) with the sampled ones (black points). The lower panels (d-f) show the mean percentage difference between the true and sampled entropy curves, with associated errors. (a) Single-layer perceptron with 15 input neurons and $P = 15$. (b) one-hidden layer network for binary classification with 5 input neurons, 3 hidden neurons and $P = 18$. (c) One-hidden layer network for multi-class classification with 3 input neurons, 3 hidden neurons, 3 output neurons and $P = 15$. The corresponding lower panels (d, e, f) illustrate the mean value of the percentage difference between the true and sampled entropy curves with the associated error bars.

F.3.2 Scaling analysis of the Wang-Landau algorithm

To understand how the WL algorithm scales with system size and problem complexity, we investigated the convergence time as a function of two key variables: the number of neurons in a single-layer perceptron and the complexity of the learning task.

Network Size Dependence

First, we analysed how convergence time varies with the number of neurons N in a single-layer perceptron, keeping the learning complexity fixed at $\alpha = \frac{P}{N} = 1$. For each N value, we generated 7 independent datasets and ran 50 WL simulations per dataset, recording average convergence times. These times, plotted in Figure F.2, exhibit a power-law dependency, approximately $T(N) \sim 10^{-6}N^5$.

Learning Complexity Dependence

Next, we examined convergence time as a function of learning complexity $\alpha = \frac{P}{N}$ with $N = 100$ fixed. For each P value, 7 datasets were generated, and 50 WL simula-

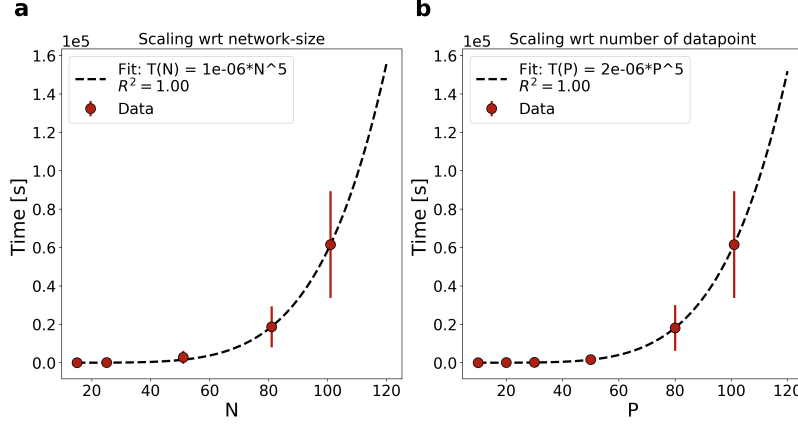


Figure F.2: Convergence time of the Wang-Landau algorithm executed on a single core as a function of network size and learning complexity. (a) The scaling of convergence time with respect to the number of neurons N in a single-layer perceptron, at a fixed learning complexity $\alpha = \frac{P}{N} = 1$. For each value of N , 7 independent datasets were generated, and for each dataset, 50 independent simulations were performed. The data points represent the mean convergence time, and the dashed line represents the fitted curve with $T(N) = 10^{-6}N^5$. (b) The convergence time as a function of the learning complexity α , at a fixed number of neurons $N = 100$. Similarly, 7 datasets were generated for each value of P , and 50 independent simulations were conducted per dataset. The fitted curve is $T(P) = 2 \times 10^{-6}P^5$.

tions were performed per dataset. As shown in Figure F.2 convergence times follow an approximate scaling of $T(P) \sim 2 \times 10^{-6}P^5$, reflecting increased computational demands as task complexity grows with larger datasets.

All simulations in this analysis were performed on a single core, which serves as a baseline for computational cost. These results could be significantly improved by leveraging multi-core processing or by implementing parallel simulations across multiple energy windows, where each window covers a specific range of the energy spectrum. This approach allows for concurrent WL sampling within each window, reducing convergence time and making the WL algorithm more suitable for large-scale or high-complexity tasks.

In summary, these scaling analyses indicate that WL sampling exhibits polynomial growth in computational cost with network size and learning complexity, both with exponents near 5. This finding provides essential insights for applying WL sampling to large neural networks or complex datasets, guiding future optimization efforts for efficient algorithm implementation.

F.4 Gaussian Fit

In this subsection, we analyse the DoS curves for binary classification using a single-layer perceptron. Our goal is to determine whether the distribution of states obtained from classifying independent and identically distributed (i.i.d.) data follows a Gaussian distribution. Figure F.3 presents the DoS curves for different numbers of input neurons, N , along with a Gaussian fit, with the fit parameters—the stan-

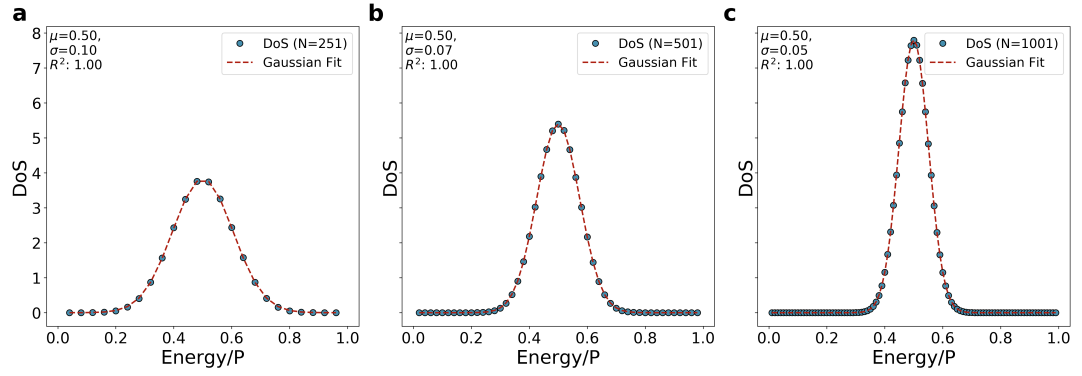


Figure F.3: *DoS* Curves for Binary Classification by a Single-Layer Perceptron and Gaussian Fit. Each panel presents the *DoS* for binary classification of random data using a single-layer perceptron with different numbers of input layer neurons, N as specified in the legend. The top left corner of each plot provides the parameters of the Gaussian fit, including the standard deviation (σ) and the mean (μ), as well as the coefficient of determination (R^2).

dard deviation (σ) and the mean (μ) of the Gaussian distribution— indicated in the upper left corner. Additionally, the goodness of fit is measured by the coefficient R^2 , also shown in the upper left corner of each panel. The R^2 value indicates how well the Gaussian model represents the observed *DoS*, with values closer to 1 signifying a better fit. Although the *DoS* is not continuous in the cases studied, the *DoS* approaches a continuous distribution in the thermodynamic limit, where both N and the number of patterns P increase indefinitely while maintaining a fixed ratio. This analysis demonstrates that, under these conditions, the *DoS* curves closely align with a Gaussian distribution, supporting the use of Gaussian models to describe the statistical behaviour of single-layer perceptrons in binary classification tasks with i.i.d. data.

F.5 Fashion-MNIST

This section provides an analysis of a subset of the FashionMNIST dataset, focusing on the binary classification of T-shirt/top and trouser images, to assess the generality of trends observed in the presence of different class imbalances. As shown in Figure F.4, the results mirror those presented in the main text, reinforcing the critical role of class imbalance in modelling DoS curves.

At lower complexity ($P/N = 0.1$), the DoS reveals a clear symmetry: as class imbalance increases, the peaks shift progressively from the centre of the spectrum, regardless of which class is predominant. This pattern is maintained even at higher complexity ($P/N = 0.5$), where the separation of the peaks becomes more pronounced with imbalance. In both cases, the location of these peaks reflects the degree of imbalance, with the largest deviations from balance showing the most extreme shifts.

The correlation between peak energy and imbalance is further highlighted in the scatter plot, where the absolute difference $|0.5 - P_1/P|$ captures the relationship between class distribution and peak shift. This trend holds across different learning complexities, demonstrating the robustness of this behaviour. These results on FashionMNIST corroborate our findings in the main text and emphasize that the DoS landscape is significantly influenced by structured data, with class imbalance driving non-trivial modifications in the energy spectrum.

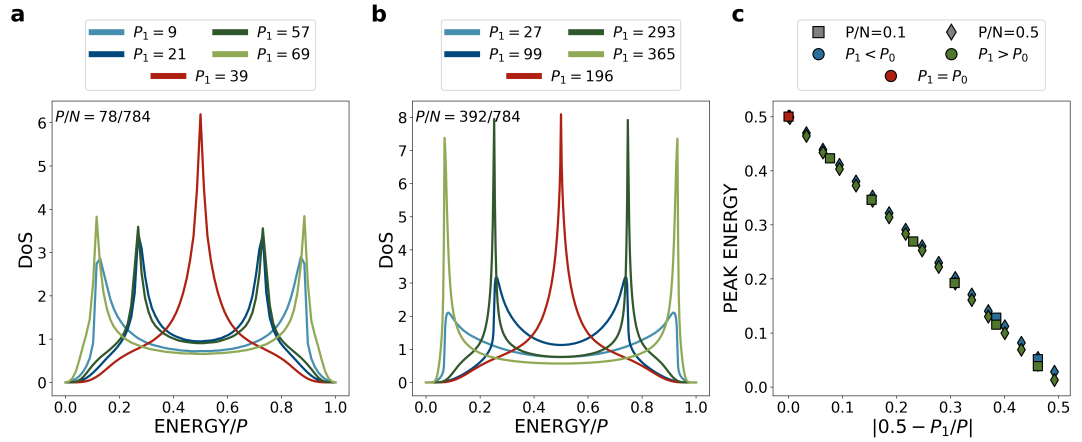


Figure F.4: DoS analysis for binary classification of FashionMNIST images (T-shirt/top and Trouser) under various class imbalances. (a) DoS at a fixed learning complexity of $P/N = 0.1$, showing the distribution for different class imbalances. The legend indicates the number of elements in class 1 (P_1). Perfect class balance is achieved when $P_1 = P_0 = P/2$ (red curve). Larger deviations from this value indicate greater class imbalance. Blue curves represent a predominance of class 0, while green curves represent a predominance of class 1; the lighter the colour, the greater the class imbalance. (b) DoS at a higher learning complexity of $P/N = 0.5$, showing similar trends with varying class imbalances. (c) Peak energy values plotted against the absolute difference $|0.5 - P_1/P|$, highlighting the correlation between peak energy and class imbalance. Blue points indicate a predominance of class 0, while green points indicate a predominance of class 1. Red points represent perfect balance ($P_1 = P_0$). Results are shown for two values of learning complexity: $P/N = 0.1$ (squares) and $P/N = 0.5$ (diamonds).

Bibliography

Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl (2015). “GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers”. In: *SoftwareX* 1, pp. 19–25. DOI: [10.1016/j.softx.2015.06.001](https://doi.org/10.1016/j.softx.2015.06.001). URL: <https://doi.org/10.1016/j.softx.2015.06.001>.

Abramyan, Tigran M., James A. Snyder, Aby A. Thyparambil, Steven J. Stuart, and Robert A. Latour (2016). “Cluster Analysis of Molecular Simulation Trajectories for Systems Where Both Conformation and Orientation of the Sampled States Are Important”. In: *Journal of Computational Chemistry* 37.21, pp. 1973–1982. DOI: [10.1002/jcc.24416](https://doi.org/10.1002/jcc.24416). URL: <https://doi.org/10.1002/jcc.24416>.

Aitchison, Laurence, Nicola Corradi, and Peter E. Latham (2016). “Zipf’s Law Arises Naturally When There Are Underlying, Unobserved Variables”. In: *PLoS Computational Biology* 12.12, e1005110. DOI: [10.1371/journal.pcbi.1005110](https://doi.org/10.1371/journal.pcbi.1005110). URL: <https://doi.org/10.1371/journal.pcbi.1005110>.

Albert, Réka and Albert-László Barabási (2002). “Statistical Mechanics of Complex Networks”. In: *Reviews of Modern Physics* 74.1, pp. 47–97. DOI: [10.1103/RevModPhys.74.47](https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47). URL: <https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47>.

Albert, Réka, Hawoong Jeong, and Albert-László Barabási (2000). “Error and Attack Tolerance of Complex Networks”. In: *Nature* 406.6794, pp. 378–382. DOI: [10.1038/35019019](https://www.nature.com/articles/35019019). URL: <https://www.nature.com/articles/35019019>.

Alder, Berni J. and Thomas E. Wainwright (1959). “Studies in Molecular Dynamics. I. General Method”. In: *The Journal of Chemical Physics* 31.2, pp. 459–466. DOI: [10.1063/1.1730376](https://doi.org/10.1063/1.1730376). URL: <https://doi.org/10.1063/1.1730376>.

Aldrigo, Riccardo, Roberto Menichetti, and Raffaello Potestio (2025). “Low-Resolution Descriptions of Model Neural Activity Reveal Hidden Features and Underlying System Properties”. In: *Physical Review E* 111.4, p. 044315. DOI: [10.1103/PhysRevE.111.044315](https://doi.org/10.1103/PhysRevE.111.044315). URL: <https://doi.org/10.1103/PhysRevE.111.044315>.

Anderson, Philip W. (1972). “More is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science”. In: *Science* 177.4047, pp. 393–396. DOI: [10.1126/science.1226699](https://doi.org/10.1126/science.1226699).

1126/science.177.4047.393. URL: <https://www.science.org/doi/10.1126/science.177.4047.393>.

Awaya, Naoki and Li Ma (2024). “Unsupervised Tree Boosting for Learning Probability Distributions”. In: *Journal of Machine Learning Research* 25.198, pp. 1–52. URL: <https://www.jmlr.org/papers/v25/22-0980.html>.

Ayon, Navid J. (2020). “Features, Roles and Chiral Analyses of Proteinogenic Amino Acids”. In: *AIMS Molecular Science* 7.3, pp. 229–268. DOI: 10.3934/molsci.2020011. URL: <https://www.aimspress.com/article/doi/10.3934/molsci.2020011>.

Bahar, Ivet, Ali Rana Atilgan, and Burak Erman (1997). “Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential”. In: *Folding and Design* 2.3, pp. 173–181. DOI: 10.1016/S1359-0278(97)00024-2. URL: [https://doi.org/10.1016/S1359-0278\(97\)00024-2](https://doi.org/10.1016/S1359-0278(97)00024-2).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.1409.0473. URL: <https://arxiv.org/abs/1409.0473>.

Bak, Per, Chao Tang, and Kurt Wiesenfeld (1987). “Self-Organized Criticality: An Explanation of the $1/f$ Noise”. In: *Physical Review Letters* 59.4, pp. 381–384. DOI: 10.1103/PhysRevLett.59.381. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.59.381>.

Baldassi, Carlo, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina (2016). “Unreasonable Effectiveness of Learning Neural Networks: From Accessible States and Robust Ensembles to Basic Algorithmic Schemes”. In: *Proceedings of the National Academy of Sciences* 113.48, E7655–E7662. DOI: 10.1073/pnas.1608103113. URL: <https://doi.org/10.1073/pnas.1608103113>.

Baldassi, Carlo, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina (2015). “Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses”. In: *Physical Review Letters* 115.12, p. 128101. DOI: 10.1103/PhysRevLett.115.128101. URL: <https://doi.org/10.1103/PhysRevLett.115.128101>.

Baldassi, Carlo, Clarissa Lauditi, Enrico M. Malatesta, Gabriele Perugini, and Riccardo Zecchina (2021). “Unveiling the Structure of Wide Flat Minima in Neural Networks”. In: *Physical Review Letters* 127.27, p. 278301. DOI: 10.1103/PhysRevLett.127.278301. URL: <https://doi.org/10.1103/PhysRevLett.127.278301>.

Ballard, Andrew J., Ritankar Das, Stefano Martiniani, Dhagash Mehta, Levent Sagun, Jacob D. Stevenson, and David J. Wales (2017). “Energy Landscapes for Ma-

chine Learning". In: *Physical Chemistry Chemical Physics* 19.20, pp. 12585–12603. DOI: 10.1039/C7CP01108C. URL: <http://dx.doi.org/10.1039/C7CP01108C>.

Bar-Yam, Yaneer (2004). "A Mathematical Theory of Strong Emergence Using Multiscale Variety". In: *Complexity* 9.6, pp. 15–24. DOI: 10.1002/cplx.20029. URL: <https://doi.org/10.1002/cplx.20029>.

Bar-Yam, Yaneer (2019). *Dynamics of Complex Systems*. CRC Press. DOI: 10.1201/9780429034961. URL: <https://doi.org/10.1201/9780429034961>.

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba (2017). "Network Dissection: Quantifying Interpretability of Deep Visual Representations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327. DOI: 10.1109/CVPR.2017.354. URL: <https://doi.org/10.1109/CVPR.2017.354>.

Beck, Thomas L., Paolo Carloni, and Dilipkumar N. Asthagiri (2024). "All-Atom Biomolecular Simulation in the Exascale Era". In: *Journal of Chemical Theory and Computation* 20.5, pp. 1777–1782. DOI: 10.1021/acs.jctc.3c01276. URL: <https://doi.org/10.1021/acs.jctc.3c01276>.

Beckstein, Oliver, Elizabeth J. Denning, Juan R. Perilla, and Thomas B. Woolf (2009). "Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open↔Closed Transitions". In: *Journal of Molecular Biology* 394.1, pp. 160–176. DOI: 10.1016/j.jmb.2009.09.009. URL: <https://doi.org/10.1016/j.jmb.2009.09.009>.

Bekker, Henk, Herman Berendsen, E.J. Dijkstra, S. Achterop, Rudi Drunen, David van der Spoel, A. Sijbers, H. Keegstra, B. Reitsma, and M.K.R. Renardus (1993). "GROMACS: A Parallel Computer for Molecular Dynamics Simulations". In: *Physics Computing '92*. Singapore: World Scientific Publishing, pp. 252–256. ISBN: 981-02-1245-3. DOI: 10.1142/1893. URL: <https://research.rug.nl/en/publications/gromacs-a-parallel-computer-for-molecular-dynamics-simulations>.

Belardinelli, Rolando E. and Victor D. Pereyra (2007a). "Fast Algorithm to Calculate Density of States". In: *Physical Review E* 75.4, p. 046701. DOI: 10.1103/PhysRevE.75.046701. URL: <https://doi.org/10.1103/PhysRevE.75.046701>.

Belardinelli, Rolando E. and Victor D. Pereyra (2007b). "Wang-Landau Algorithm: A Theoretical Analysis of the Saturation of the Error". In: *The Journal of Chemical Physics* 127.18, p. 184105. DOI: 10.1063/1.2803061. URL: <https://doi.org/10.1063/1.2803061>.

Belardinelli, Rolando E., Victor D. Pereyra, Ronald Dickman, and Bruno Jeferson Lourenço (2014). "Intrinsic Convergence Properties of Entropic Sampling Algorithms". In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.7, P07007. DOI:

10.1088/1742-5468/2014/07/P07007. URL: <https://doi.org/10.1088/1742-5468/2014/07/P07007>.

Bellman, Richard E. (2015). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press. DOI: 10.1515/9781400874668. URL: <https://doi.org/10.1515/9781400874668>.

Belrose, Nora, Quintin Pope, Lucia Quirke, Alex Troy Mallen, and Xiaoli Fern (2024). “Neural Networks Learn Statistics of Increasing Complexity”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 3382–3409. URL: <https://proceedings.mlr.press/v235/belrose24a.html>.

Bengio, Yoshua (2012). “Practical Recommendations for Gradient-Based Training of Deep Architectures”. In: *Neural Networks: Tricks of the Trade*. Springer, pp. 437–478. DOI: 10.1007/978-3-642-35289-8_26. URL: https://doi.org/10.1007/978-3-642-35289-8_26.

Berg, Bernd A. and Thomas Neuhaus (1992). “Multicanonical Ensemble: A New Approach to Simulate First-Order Phase Transitions”. In: *Physical Review Letters* 68.1, pp. 9–12. DOI: 10.1103/PhysRevLett.68.9. URL: <https://doi.org/10.1103/PhysRevLett.68.9>.

Best, Robert B. and Gerhard Hummer (2005). “Reaction Coordinates and Rates from Transition Paths”. In: *Proceedings of the National Academy of Sciences* 102.19, pp. 6732–6737. DOI: 10.1073/pnas.0408098102. URL: <https://doi.org/10.1073/pnas.0408098102>.

Bianconi, Ginestra, Alex Arenas, Jacob Biamonte, Lincoln D. Carr, Byungnam Kahng, Janos Kertész, Jürgen Kurths, Linyuan Lü, Cristina Masoller, Adilson E. Motter, Matjaž Perc, Filippo Radicchi, Ramakrishna Ramaswamy, Francisco A. Rodrigues, Marta Sales-Pardo, Maxi San Miguel, Stefan Thurner, and Taha Yasseri (2023). “Complex Systems in the Spotlight: Next Steps after the 2021 Nobel Prize in Physics”. In: *Journal of Physics: Complexity* 4.1, p. 010201. DOI: 10.1088/2632-072X/ac7f75. URL: <https://iopscience.iop.org/article/10.1088/2632-072X/ac7f75>.

Biehl, Michael and Holm Schwarze (1995). “Learning by On-Line Gradient Descent”. In: *Journal of Physics A: Mathematical and General* 28.3, pp. 643–656. DOI: 10.1088/0305-4470/28/3/018. URL: <https://doi.org/10.1088/0305-4470/28/3/018>.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer. URL: <https://link.springer.com/book/10.1007/978-0-387-45528-0>.

Blalock, Davis, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Gutttag (2020). "What Is the State of Neural Network Pruning?" In: *Proceedings of Machine Learning and Systems* 2, pp. 129–146. URL: https://proceedings.mlsys.org/paper_files/paper/2020/hash/6c44dc73014d66ba49b28d483a8f8b0d-Abstract.html.

Boehr, David D., Ruth Nussinov, and Peter E. Wright (2009). "The Role of Dynamic Conformational Ensembles in Biomolecular Recognition". In: *Nature Chemical Biology* 5.11, pp. 789–796. DOI: 10.1038/nchembio.232. URL: <https://doi.org/10.1038/nchembio.232>.

Bolhuis, Peter G., David Chandler, Christoph Dellago, and Phillip L. Geissler (2002). "Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark". In: *Annual Review of Physical Chemistry* 53.1, pp. 291–318. DOI: 10.1146/annurev.physchem.53.082301.113146. URL: <https://doi.org/10.1146/annurev.physchem.53.082301.113146>.

Bolhuis, Peter G., Christoph Dellago, and David Chandler (2000). "Reaction Coordinates of Biomolecular Isomerization". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.11, pp. 5877–5882. DOI: 10.1073/pnas.100127697. URL: <https://doi.org/10.1073/pnas.100127697>.

Bonachela, Juan A., Haye Hinrichsen, and Miguel A. Muñoz (2008). "Entropy Estimates of Small Data Sets". In: *Journal of Physics A: Mathematical and Theoretical* 41.20, p. 202001. DOI: 10.1088/1751-8113/41/20/202001. URL: <https://doi.org/10.1088/1751-8113/41/20/202001>.

Bottou, Léon (2012). "Stochastic Gradient Descent Tricks". In: *Neural Networks: Tricks of the Trade*. Springer, pp. 421–436. DOI: 10.1007/978-3-642-35289-8_25. URL: https://doi.org/10.1007/978-3-642-35289-8_25.

Bryngelson, Joseph D. and Peter G. Wolynes (1987). "Spin Glasses and the Statistical Mechanics of Protein Folding". In: *Proceedings of the National Academy of Sciences of the United States of America* 84.21, pp. 7524–7528. DOI: 10.1073/pnas.84.21.7524. URL: <https://www.pnas.org/doi/10.1073/pnas.84.21.7524>.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello (2007). "Canonical Sampling through Velocity Rescaling". In: *The Journal of Chemical Physics* 126.1, p. 014101. DOI: 10.1063/1.2408420. URL: <https://doi.org/10.1063/1.2408420>.

Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naf-tali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová (2019). "Machine Learning and the Physical Sciences". In: *Reviews of Modern Physics* 91.4, p. 045002. DOI: 10.1103/RevModPhys.91.045002. URL: <https://doi.org/10.1103/RevModPhys.91.045002>.

Carlton, A. George (1969). "On the Bias of Information Estimates". In: *Psychological Bulletin* 71.2, pp. 108–109. DOI: 10.1037/h0026857. URL: <https://doi.org/10.1037/h0026857>.

Chaimovich, Aviel and M. Scott Shell (2011). "Coarse-Graining Errors and Numerical Optimization Using a Relative Entropy Framework". In: *The Journal of Chemical Physics* 134.9, p. 094112. DOI: 10.1063/1.3557038. URL: <https://doi.org/10.1063/1.3557038>.

Cheng, Hongrong, Miao Zhang, and Javen Qinfeng Shi (2024). "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12, pp. 10558–10578. DOI: 10.1109/TPAMI.2024.3447085. URL: <https://doi.org/10.1109/TPAMI.2024.3447085>.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: <https://doi.org/10.3115/v1/D14-1179>.

Chodera, John D. and Frank Noé (2014). "Markov State Models of Biomolecular Conformational Dynamics". In: *Current Opinion in Structural Biology* 25, pp. 135–144. DOI: 10.1016/j.sbi.2014.04.002. URL: <https://doi.org/10.1016/j.sbi.2014.04.002>.

Chodera, John D., Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope (2007). "Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics". In: *The Journal of Chemical Physics* 126.15, p. 155101. DOI: 10.1063/1.2714538. URL: <https://doi.org/10.1063/1.2714538>.

Choromanska, Anna, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun (2015). "The Loss Surfaces of Multilayer Networks". In: *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*. Vol. 38. PMLR. Proceedings of Machine Learning Research, pp. 192–204. URL: <https://proceedings.mlr.press/v38/choromanska15.html>.

Coifman, Ronald R., Stephane Lafon, Ann B. Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W. Zucker (2005). "Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21, pp. 7426–7431. DOI: 10.1073/pnas.0500334102. URL: <https://www.pnas.org/doi/10.1073/pnas.0500334102>.

Cornacchia, Elisabetta, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová (2023). “Learning Curves for the Multi-Class Teacher–Student Perceptron”. In: *Machine Learning: Science and Technology* 4.1, p. 015019. DOI: [10.1088/2632-2153/acb428](https://doi.org/10.1088/2632-2153/acb428). URL: <https://doi.org/10.1088/2632-2153/acb428>.

Cornell, Wendy D., Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman (1995). “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. In: *Journal of the American Chemical Society* 117.19, pp. 5179–5197. DOI: [10.1021/ja00124a002](https://doi.org/10.1021/ja00124a002). URL: <https://doi.org/10.1021/ja00124a002>.

Costa, Madalena, Ary L. Goldberger, and C.-K. Peng (2005). “Multiscale Entropy Analysis of Biological Signals”. In: *Physical Review E* 71.2, p. 021906. DOI: [10.1103/PhysRevE.71.021906](https://doi.org/10.1103/PhysRevE.71.021906). URL: <https://doi.org/10.1103/PhysRevE.71.021906>.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. 2nd. John Wiley & Sons. DOI: [10.1002/047174882X](https://doi.org/10.1002/047174882X). URL: <https://doi.org/10.1002/047174882X>.

Crick, Francis (1970). “Central Dogma of Molecular Biology”. In: *Nature* 227.5258, pp. 561–563. DOI: [10.1038/227561a0](https://www.nature.com/articles/227561a0). URL: <https://www.nature.com/articles/227561a0>.

Cubero, Ryan John, Junghyo Jo, Matteo Marsili, Yasser Roudi, and Juyong Song (2019). “Statistical Criticality Arises in Most Informative Representations”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.6, p. 063402. DOI: [10.1088/1742-5468/ab16c8](https://doi.org/10.1088/1742-5468/ab16c8). URL: <https://doi.org/10.1088/1742-5468/ab16c8>.

Cubero, Ryan John, Matteo Marsili, and Yasser Roudi (2020). “Multiscale Relevance and Informative Encoding in Neuronal Spike Trains”. In: *Journal of Computational Neuroscience* 48.1, pp. 85–102. DOI: [10.1007/s10827-020-00740-x](https://doi.org/10.1007/s10827-020-00740-x). URL: <https://doi.org/10.1007/s10827-020-00740-x>.

Cybenko, George (1989). “Approximation by Superpositions of a Sigmoidal Function”. In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314. DOI: [10.1007/BF02551274](https://doi.org/10.1007/BF02551274). URL: <https://doi.org/10.1007/BF02551274>.

Darden, Tom, Darrin York, and Lee Pedersen (1993). “Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems”. In: *The Journal of Chemical Physics* 98.12, pp. 10089–10092. DOI: [10.1063/1.464397](https://doi.org/10.1063/1.464397). URL: <https://doi.org/10.1063/1.464397>.

Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli (2021). “Ad-

vancing Mathematics by Guiding Human Intuition with AI". In: *Nature* 600.7887, pp. 70–74. DOI: 10.1038/s41586-021-04086-x. URL: <https://doi.org/10.1038/s41586-021-04086-x>.

Davtyan, Aram, Nicholas P. Schafer, Weihua Zheng, Cecilia Clementi, Peter G. Wolynes, and Garegin A. Papoian (2012). "AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing". In: *The Journal of Physical Chemistry B* 116.29, pp. 8494–8503. DOI: 10.1021/jp212541y. URL: <https://doi.org/10.1021/jp212541y>.

De la Porte, J., Ben M. Herbst, Willy Hereman, and Steèfan Johann van der Walt (2008). "An Introduction to Diffusion Maps". In: *Proceedings of the 19th Symposium of the Pattern Recognition Association of South Africa (PRASA 2008)*. Cape Town, South Africa, pp. 15–25. URL: <https://citeseerx.ist.psu.edu/doc/10.1.1.309.674>.

Dellago, Christoph, Peter G. Bolhuis, and Phillip L. Geissler (2002). "Transition Path Sampling". In: *Advances in Chemical Physics* 123, pp. 1–78. DOI: 10.1002/0471231509.ch1. URL: <https://doi.org/10.1002/0471231509.ch1>.

Deng, Li (Nov. 2012). "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142. DOI: 10.1109/MSP.2012.2211477. URL: <https://doi.org/10.1109/MSP.2012.2211477>.

Diggins IV, Patrick, Changjiang Liu, Markus Deserno, and Raffaello Potestio (2019). "Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules". In: *Journal of Chemical Theory and Computation* 15.1, pp. 648–664. DOI: 10.1021/acs.jctc.8b00654. URL: <https://doi.org/10.1021/acs.jctc.8b00654>.

Dill, Ken A. and Hue Sun Chan (1997). "From Levinthal to Pathways to Funnels". In: *Nature Structural Biology* 4.1, pp. 10–19. DOI: 10.1038/nsb0197-10. URL: <https://www.nature.com/articles/nsb0197-10>.

Dill, Ken A., S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl (2008). "The Protein Folding Problem". In: *Annual Review of Biophysics* 37.1, pp. 289–316. DOI: 10.1146/annurev.biophys.37.092707.153558. URL: <https://doi.org/10.1146/annurev.biophys.37.092707.153558>.

Dror, Ron O., Robert M. Dirks, J. P. Grossman, Huafeng Xu, and David E. Shaw (2012). "Biomolecular Simulation: A Computational Microscope for Molecular Biology". In: *Annual Review of Biophysics* 41, pp. 429–452. DOI: 10.1146/annurev-biophys-042910-155245. URL: <https://doi.org/10.1146/annurev-biophys-042910-155245>.

Duranthon, O., Matteo Marsili, and Rongrong Xie (2021). "Maximal Relevance and Optimal Learning Machines". In: *Journal of Statistical Mechanics: Theory and Experi-*

ment 2021.3, p. 033409. DOI: [10.1088/1742-5468/abe6ff](https://doi.org/10.1088/1742-5468/abe6ff). URL: <https://doi.org/10.1088/1742-5468/abe6ff>.

Eastman, Peter, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Benjamin R. Brooks, and Vijay S. Pande (2017). "OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics". In: *PLoS Computational Biology* 13.7, e1005659. DOI: [10.1371/journal.pcbi.1005659](https://doi.org/10.1371/journal.pcbi.1005659). URL: <https://doi.org/10.1371/journal.pcbi.1005659>.

Elman, Jeffrey L. (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211. ISSN: 0364-0213. DOI: [10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E). URL: [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E).

Engel, Andreas and Christian van den Broeck (2001). *Statistical Mechanics of Learning*. Cambridge: Cambridge University Press. DOI: [10.1017/CB09781139164542](https://doi.org/10.1017/CB09781139164542). URL: <https://doi.org/10.1017/CB09781139164542>.

Engel, Eberhard (2011). *Density Functional Theory*. Springer. DOI: [10.1007/978-3-642-14090-7](https://doi.org/10.1007/978-3-642-14090-7). URL: <https://doi.org/10.1007/978-3-642-14090-7>.

Essmann, Ulrich, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen (1995). "A Smooth Particle Mesh Ewald Method". In: *The Journal of Chemical Physics* 103.19, pp. 8577–8593. DOI: [10.1063/1.470117](https://doi.org/10.1063/1.470117). URL: <https://doi.org/10.1063/1.470117>.

Farnia, Farzan, Jesse M. Zhang, and David N. C. Tse (2020). "A Fourier-Based Approach to Generalization and Optimization in Deep Learning". In: *IEEE Journal on Selected Areas in Information Theory* 1.1, pp. 145–156. DOI: [10.1109/JSAIT.2020.2983192](https://doi.org/10.1109/JSAIT.2020.2983192). URL: <https://doi.org/10.1109/JSAIT.2020.2983192>.

Farris, Alfred C. K. and David P. Landau (2021). "Replica Exchange Wang–Landau Sampling of Long HP Model Sequences". In: *Physica A: Statistical Mechanics and its Applications* 569, p. 125778. DOI: [10.1016/j.physa.2021.125778](https://doi.org/10.1016/j.physa.2021.125778). URL: <https://doi.org/10.1016/j.physa.2021.125778>.

Fisher, Michael E. (1998). "Renormalization Group Theory: Its Basis and Formulation in Statistical Physics". In: *Reviews of Modern Physics* 70.2, pp. 653–681. DOI: [10.1103/RevModPhys.70.653](https://doi.org/10.1103/RevModPhys.70.653). URL: <https://doi.org/10.1103/RevModPhys.70.653>.

Foley, Thomas T., Katherine M. Kidder, M. Scott Shell, and William G. Noid (2020). "Exploring the Landscape of Model Representations". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.39, pp. 24061–24068. DOI: [10.1073/pnas.2000098117](https://doi.org/10.1073/pnas.2000098117). URL: <https://doi.org/10.1073/pnas.2000098117>.

Foley, Thomas T., M. Scott Shell, and William George Noid (2015). "The Impact of Resolution upon Entropy and Information in Coarse-Grained Models". In: *The*

Journal of Chemical Physics 143.24, p. 243104. DOI: 10.1063/1.4929836. URL: <https://doi.org/10.1063/1.4929836>.

Francazi, Emanuele, Marco Baity-Jesi, and Aurelien Lucchi (2023). “A Theoretical Analysis of the Learning Dynamics under Class Imbalance”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 10285–10322. URL: <https://proceedings.mlr.press/v202/francazi23a.html>.

Francazi, Emanuele, Aurelien Lucchi, and Marco Baity-Jesi (2024). “Initial Guessing Bias: How Untrained Networks Favor Some Classes”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 13783–13839. URL: <https://proceedings.mlr.press/v235/francazi24a.html>.

Frankle, Jonathan and Michael Carbin (2019). “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=rJl-b3RcF7>.

Frauenfelder, Hans, Stephen G. Sligar, and Peter G. Wolynes (1991). “The Energy Landscapes and Motions of Proteins”. In: *Science* 254.5038, pp. 1598–1603. DOI: 10.1126/science.1749933. URL: <https://www.science.org/doi/10.1126/science.1749933>.

Freddolino, Peter L., Anton S. Arkhipov, Steven B. Larson, Alexander McPherson, and Klaus Schulten (2006). “Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus”. In: *Structure* 14.3, pp. 437–449. DOI: 10.1016/j.str.2005.11.014. URL: <https://doi.org/10.1016/j.str.2005.11.014>.

Freedman, David and Persi Diaconis (1981). “On the Histogram as a Density Estimator: L_2 Theory”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57.4, pp. 453–476. DOI: 10.1007/BF01025868. URL: <https://doi.org/10.1007/BF01025868>.

Frenkel, Daan and Berend Smit (2023). *Understanding Molecular Simulation: From Algorithms to Applications*. 3rd. Elsevier. DOI: 10.1016/B978-012267351-1/50033-X. URL: <https://www.sciencedirect.com/science/article/pii/B978012267351150033X>.

Funahashi, Ken-Ichi (1989). “On the Approximate Realization of Continuous Mappings by Neural Networks”. In: *Neural Networks* 2.3, pp. 183–192. DOI: 10.1016/0893-6080(89)90003-8. URL: [https://doi.org/10.1016/0893-6080\(89\)90003-8](https://doi.org/10.1016/0893-6080(89)90003-8).

Fushing, Hsieh, Hui Wang, Kimberly VanderWaal, Brenda McCowan, and Patrice Koehl (2013). “Multi-Scale Clustering by Building a Robust and Self Correcting

Ultrametric Topology on Data Points". In: *PLOS ONE* 8.2, e56259. DOI: 10.1371/journal.pone.0056259. URL: <https://doi.org/10.1371/journal.pone.0056259>.

Gapsys, Vytautas, Wojciech Kopec, Dirk Matthes, and Bert L. de Groot (2024). "Biomolecular Simulations at the Exascale: From Drug Design to Organelles and Beyond". In: *Current Opinion in Structural Biology* 88, p. 102887. DOI: 10.1016/j.sbi.2024.102887. URL: <https://doi.org/10.1016/j.sbi.2024.102887>.

Gardner, Elizabeth (1988). "The Space of Interactions in Neural Network Models". In: *Journal of Physics A: Mathematical and General* 21.1, pp. 257–270. DOI: 10.1088/0305-4470/21/1/030. URL: <https://doi.org/10.1088/0305-4470/21/1/030>.

Gardner, Elizabeth and Bernard Derrida (1988). "Optimal Storage Properties of Neural Network Models". In: *Journal of Physics A: Mathematical and General* 21.1, pp. 271–284. DOI: 10.1088/0305-4470/21/1/031. URL: <https://doi.org/10.1088/0305-4470/21/1/031>.

Gerace, Federica, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová (2024). "Gaussian Universality of Perceptrons with Random Labels". In: *Physical Review E* 109.3, p. 034305. DOI: 10.1103/PhysRevE.109.034305. URL: <https://doi.org/10.1103/PhysRevE.109.034305>.

Gerace, Federica, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová (2021). "Generalisation Error in Learning with Random Features and the Hidden Manifold Model". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12, p. 124013. DOI: 10.1088/1742-5468/ac3ae6. URL: <https://doi.org/10.1088/1742-5468/ac3ae6>.

Giulini, Marco (2022). "The Mapping Problem in Coarse-Grained Modelling of Biomolecules". PhD thesis. Università degli Studi di Trento. URL: https://iris.unitn.it/retrieve/handle/11572/330532/524536/phd_unitn_marco_giulini.pdf.

Giulini, Marco, Raffaele Fiorentini, Luca Tubiana, Raffaello Potestio, and Roberto Menichetti (2024). "EXCOGITO, an Extensible Coarse-Graining Toolbox for the Investigation of Biomolecules by Means of Low-Resolution Representations". In: *Journal of Chemical Information and Modeling* 64.12, pp. 4912–4927. DOI: 10.1021/acs.jcim.4c00490. URL: <https://doi.org/10.1021/acs.jcim.4c00490>.

Giulini, Marco, Roberto Menichetti, M. Scott Shell, and Raffaello Potestio (2020). "An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules". In: *Journal of Chemical Theory and Computation* 16.11, pp. 6795–6813. DOI: 10.1021/acs.jctc.0c00676. URL: <https://doi.org/10.1021/acs.jctc.0c00676>.

Giulini, Marco, Marta Rigoli, Giovanni Mattiotti, Roberto Menichetti, Thomas Tarenzi, Raffaele Fiorentini, and Raffaello Potestio (2021). "From System Modeling

to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules". In: *Frontiers in Molecular Biosciences* 8, p. 676976. DOI: 10.3389/fmolb.2021.676976. URL: <https://doi.org/10.3389/fmolb.2021.676976>.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 315–323. URL: <https://proceedings.mlr.press/v15/glorot11a.html>.

Goldenfeld, Nigel (2018). *Lectures on Phase Transitions and the Renormalization Group*. CRC Press. DOI: 10.1201/9780429493492. URL: <https://doi.org/10.1201/9780429493492>.

Goldenfeld, Nigel and Leo P. Kadanoff (1999). "Simple Lessons from Complexity". In: *Science* 284.5411, pp. 87–89. DOI: 10.1126/science.284.5411.87. URL: <https://www.science.org/doi/10.1126/science.284.5411.87>.

Goldt, Sebastian, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová (2019). "Dynamics of Stochastic Gradient Descent for Two-Layer Neural Networks in the Teacher-Student Setup". In: *Advances in Neural Information Processing Systems* 32, pp. 6979–6989. URL: <https://papers.nips.cc/paper/8921-dynamics-of-stochastic-gradient-descent-for-two-layer-neural-networks-in-the-teacher-student-setup>.

Goldt, Sebastian, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová (2020). "Dynamics of Stochastic Gradient Descent for Two-Layer Neural Networks in the Teacher-Student Setup". In: *Journal of Statistical Mechanics: Theory and Experiment* 12, p. 124010. DOI: 10.1088/1742-5468/abc61e. URL: <https://doi.org/10.1088/1742-5468/abc61e>.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press. URL: <https://www.deeplearningbook.org/>.

Grigolon, Silvia, Silvio Franz, and Matteo Marsili (2016). "Identifying Relevant Positions in Proteins by Critical Variable Selection". In: *Molecular BioSystems* 12.7, pp. 2147–2158. DOI: 10.1039/C6MB00047A. URL: <https://doi.org/10.1039/C6MB00047A>.

Grossfield, Alan, Paul N. Patrone, Daniel R. Roe, Andrew J. Schultz, Daniel W. Siderius, and Daniel M. Zuckerman (2018). "Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations". In: *Living Journal of Computational Molecular Science* 1.1, p. 5067. DOI: 10.33011/livecoms.1.1.5067. URL: <https://doi.org/10.33011/livecoms.1.1.5067>.

Grünwald, Peter D (2007). *The minimum description length principle*. MIT press.

- Haimovici, Ariel and Matteo Marsili (2015). "Criticality of Mostly Informative Samples: A Bayesian Model Selection Approach". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.10, P10013. doi: 10.1088/1742-5468/2015/10/P10013. URL: <https://doi.org/10.1088/1742-5468/2015/10/P10013>.
- Han, Song, Huizi Mao, and William J. Dally (2016). "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding". In: *International Conference on Learning Representations (ICLR)*. doi: 10.48550/arXiv.1510.00149. URL: <https://arxiv.org/abs/1510.00149>.
- Hänggi, Peter, Peter Talkner, and Michal Borkovec (1990). "Reaction-Rate Theory: Fifty Years After Kramers". In: *Reviews of Modern Physics* 62.2, pp. 251–341. doi: 10.1103/RevModPhys.62.251. URL: <https://doi.org/10.1103/RevModPhys.62.251>.
- Hassibi, Babak, David G. Stork, and Gregory J. Wolff (1993). "Optimal Brain Surgeon and General Network Pruning". In: *Proceedings of the IEEE International Conference on Neural Networks*. Vol. 1. IEEE, pp. 293–299. doi: 10.1109/ICNN.1993.298572. URL: <https://doi.org/10.1109/ICNN.1993.298572>.
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. doi: 10.1007/978-0-387-84858-7. URL: <https://doi.org/10.1007/978-0-387-84858-7>.
- Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1, pp. 97–109. doi: 10.1093/biomet/57.1.97. URL: <https://doi.org/10.1093/biomet/57.1.97>.
- Hénin, Jérôme, Tony Lelièvre, Michael R. Shirts, Omar Valsson, and Lucie Delemotte (2022). "Enhanced Sampling Methods for Molecular Dynamics Simulations". In: *Living Journal of Computational Molecular Science* 4.1, p. 1583. doi: 10.33011/livecoms.4.1.1583. URL: <https://doi.org/10.33011/livecoms.4.1.1583>.
- Henzler-Wildman, Katherine A., Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern (2007). "A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis". In: *Nature* 450.7171, pp. 913–916. doi: 10.1038/nature06407. URL: <https://doi.org/10.1038/nature06407>.
- Hess, Berk, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije (1997). "LINCS: A Linear Constraint Solver for Molecular Simulations". In: *Journal of Computational Chemistry* 18.12, pp. 1463–1472. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H. URL: [https://doi.org/10.1002/\(SICI\)1096-987X\(199709\)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H).
- Hinton, Geoffrey E. and Ruslan R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786, pp. 504–507. doi: 10.1126/science.1127647. URL: <https://doi.org/10.1126/science.1127647>.

Hinton, Geoffrey E., Oriol Vinyals, and Jeffrey Dean (2015). “Distilling the Knowledge in a Neural Network”. In: *CoRR* abs/1503.02531. doi: 10.48550/arXiv.1503.02531. URL: <https://arxiv.org/abs/1503.02531>.

Hochreiter, Sepp and Jürgen Schmidhuber (1997a). “Flat Minima”. In: *Neural Computation* 9.1, pp. 1–42. doi: 10.1162/neco.1997.9.1.1. URL: <https://doi.org/10.1162/neco.1997.9.1.1>.

Hochreiter, Sepp and Jürgen Schmidhuber (1997b). “Flat Minima”. In: *Neural Computation* 9.1, pp. 1–42. doi: 10.1162/neco.1997.9.1.1. URL: <https://doi.org/10.1162/neco.1997.9.1.1>.

Hochreiter, Sepp and Jürgen Schmidhuber (1997c). “Long Short-Term Memory”. In: *Neural Computation* 9.8, pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>.

Hockney, Roger W. and James W. Eastwood (2021). *Computer Simulation Using Particles*. CRC Press. doi: 10.1201/9780367806934. URL: <https://doi.org/10.1201/9780367806934>.

Hollingsworth, Scott A. and Ron O. Dror (2018). “Molecular Dynamics Simulation for All”. In: *Neuron* 99.6, pp. 1129–1143. doi: 10.1016/j.neuron.2018.08.011. URL: <https://doi.org/10.1016/j.neuron.2018.08.011>.

Holtzman, Roi, Marco Giulini, and Raffaello Potestio (2022). “Making Sense of Complex Systems through Resolution, Relevance, and Mapping Entropy”. In: *Physical Review E* 106.4, p. 044101. doi: 10.1103/PhysRevE.106.044101. URL: <https://doi.org/10.1103/PhysRevE.106.044101>.

Hooker, Sara, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim (2019). “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: *Advances in Neural Information Processing Systems* 32, pp. 9737–9748. doi: 10.5555/3454287.3455160. URL: <https://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks>.

Hopfield, John J. (1982). “Neural Networks and Physical Systems with Emergent Collective Computational Abilities”. In: *Proceedings of the National Academy of Sciences* 79.8, pp. 2554–2558. doi: 10.1073/pnas.79.8.2554. URL: <https://doi.org/10.1073/pnas.79.8.2554>.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). “Multilayer Feed-forward Networks Are Universal Approximators”. In: *Neural Networks* 2.5, pp. 359–366. doi: 10.1016/0893-6080(89)90020-8. URL: [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).

- Husic, Brooke E. and Vijay S. Pande (2018). "Markov State Models: From an Art to a Science". In: *Journal of the American Chemical Society* 140.7, pp. 2386–2396. DOI: 10.1021/jacs.7b12191. URL: <https://doi.org/10.1021/jacs.7b12191>.
- Ingrosso, Alessandro (2020). "Optimal Learning with Excitatory and Inhibitory Synapses". In: *PLOS Computational Biology* 16.12, e1008536. DOI: 10.1371/journal.pcbi.1008536. URL: <https://doi.org/10.1371/journal.pcbi.1008536>.
- Ingrosso, Alessandro and Sebastian Goldt (2022). "Data-Driven Emergence of Convolutional Structure in Neural Networks". In: *Proceedings of the National Academy of Sciences* 119.40, e2201854119. DOI: 10.1073/pnas.2201854119. URL: <https://doi.org/10.1073/pnas.2201854119>.
- Iofinova, Eugenia, Alexandra Peste, Mark Kurtz, and Dan Alistarh (2022). "How Well Do Sparse ImageNet Models Transfer?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12266–12276. DOI: 10.1109/CVPR52688.2022.01195. URL: <https://doi.org/10.1109/CVPR52688.2022.01195>.
- Izvekov, Sergei and Gregory A. Voth (2005). "Multiscale Coarse Graining of Liquid-State Systems". In: *The Journal of Chemical Physics* 123.13, p. 134105. DOI: 10.1063/1.2038787. URL: <https://doi.org/10.1063/1.2038787>.
- Jardine, Nicholas and Robin Sibson (1968). "A Model for Taxonomy". In: *Mathematical Biosciences* 2.3–4, pp. 465–482. DOI: 10.1016/0025-5564(68)90030-8. URL: [https://doi.org/10.1016/0025-5564\(68\)90030-8](https://doi.org/10.1016/0025-5564(68)90030-8).
- Johnson, Stephen C. (1967). "Hierarchical Clustering Schemes". In: *Psychometrika* 32.3, pp. 241–254. DOI: 10.1007/BF02289588. URL: <https://doi.org/10.1007/BF02289588>.
- Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein (1983). "Comparison of Simple Potential Functions for Simulating Liquid Water". In: *The Journal of Chemical Physics* 79.2, pp. 926–935. DOI: 10.1063/1.445869. URL: <https://doi.org/10.1063/1.445869>.
- Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features". In: *Biopolymers* 22.12, pp. 2577–2637. DOI: 10.1002/bip.360221211. URL: <https://doi.org/10.1002/bip.360221211>.
- Kadanoff, Leo P. (1966). "Scaling Laws for Ising Models Near T_c ". In: *Physique Physique Fizika* 2.6, pp. 263–272. DOI: 10.1103/PhysicsPhysiqueFizika.2.263. URL: <https://doi.org/10.1103/PhysicsPhysiqueFizika.2.263>.
- Kalimeris, Dimitris, Gal Kaplun, Preetum Nakkiran, Ben Edelman, Tristan Yang, Boaz Barak, and Fred Zhang (2019). "SGD on Neural Networks Learns Functions

of Increasing Complexity". In: *Advances in Neural Information Processing Systems* 32, pp. 3491–3501. URL: <https://proceedings.neurips.cc/paper/2019/hash/b432f34c5a997c8e7c806a895ecc5e25-Abstract.html>.

Karplus, Martin and James Andrew McCammon (2002). "Molecular Dynamics Simulations of Biomolecules". In: *Nature Structural Biology* 9.9, pp. 646–652. doi: 10.1038/nsb0902-646. URL: <https://www.nature.com/articles/nsb0902-646>.

Keskar, Nitish Shirish, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang (2017). "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=H1oyRlYgg>.

Killian, Benjamin J., Joslyn Yundenfreund Kravitz, and Michael K. Gilson (2007). "Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation". In: *The Journal of Chemical Physics* 127.2, p. 024107. doi: 10.1063/1.2746329. URL: <https://doi.org/10.1063/1.2746329>.

Kim, Jaegil, John Edward Straub, and Thomas Keyes (2006). "Statistical-Temperature Monte Carlo and Molecular Dynamics Algorithms". In: *Physical Review Letters* 97.5, p. 050601. doi: 10.1103/PhysRevLett.97.050601. URL: <https://doi.org/10.1103/PhysRevLett.97.050601>.

Kim, Jaegil, John Edward Straub, and Thomas Keyes (2012). "Replica Exchange Statistical Temperature Molecular Dynamics Algorithm". In: *The Journal of Physical Chemistry B* 116.29, pp. 8646–8653. doi: 10.1021/jp300366j. URL: <https://doi.org/10.1021/jp300366j>.

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=33X9fd2-9FyZd>.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). "Optimization by Simulated Annealing". In: *Science* 220.4598, pp. 671–680. doi: 10.1126/science.220.4598.671. URL: <https://doi.org/10.1126/science.220.4598.671>.

Kitao, Akio, Steven Hayward, and Nobuhiro Go (1998). "Energy Landscape of a Native Protein: Jumping-among-Minima Model". In: *Proteins: Structure, Function, and Bioinformatics* 33.4, pp. 496–517. doi: 10.1002/(SICI)1097-0134(19981201)33:4<496::AID-PROT4>3.0.CO;2-1. URL: [https://doi.org/10.1002/\(SICI\)1097-0134\(19981201\)33:4%3C496::AID-PROT4%3E3.0.CO;2-1](https://doi.org/10.1002/(SICI)1097-0134(19981201)33:4%3C496::AID-PROT4%3E3.0.CO;2-1).

Klein, Florencia, Martín Soñora, Lucianna Helene Santos, Ezequiel Nazareno Frigini, Andrés Ballesteros-Casallas, Matías Rodrigo Machado, and Sergio Pantano (2023). "The SIRAH Force Field: A Suite for Simulations of Complex Biological Systems at the Coarse-Grained and Multiscale Levels". In: *Journal of Structural Biology*

215.3, p. 107985. doi: 10.1016/j.jsb.2023.107985. URL: <https://doi.org/10.1016/j.jsb.2023.107985>.

Klimovich, Pavel V., Michael R. Shirts, and David L. Mobley (2015). "Guidelines for the Analysis of Free Energy Calculations". In: *Journal of Computer-Aided Molecular Design* 29.5, pp. 397–411. doi: 10.1007/s10822-015-9840-9. URL: <https://doi.org/10.1007/s10822-015-9840-9>.

Kmiecik, Sebastian, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski (2016). "Coarse-Grained Protein Models and Their Applications". In: *Chemical Reviews* 116.14, pp. 7898–7936. doi: 10.1021/acs.chemrev.6b00163. URL: <https://doi.org/10.1021/acs.chemrev.6b00163>.

Kullback, Solomon and Richard A. Leibler (1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. doi: 10.1214/aoms/1177729694. URL: <https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full>.

Ladyman, James, James Lambert, and Karoline Wiesner (2013). "What Is a Complex System?" In: *European Journal for Philosophy of Science* 3.1, pp. 33–67. doi: 10.1007/s13194-012-0056-8. URL: <https://doi.org/10.1007/s13194-012-0056-8>.

Lafon, Stéphane and Ann B. Lee (2006). "Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.9, pp. 1393–1403. doi: 10.1109/TPAMI.2006.184. URL: <https://doi.org/10.1109/TPAMI.2006.184>.

Laio, Alessandro and Michele Parrinello (2002). "Escaping Free-Energy Minima". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.20, pp. 12562–12566. doi: 10.1073/pnas.202427399. URL: <https://doi.org/10.1073/pnas.202427399>.

Landau, David P., Shan-Ho Tsai, and M. Exler (2004). "A New Approach to Monte Carlo Simulations in Statistical Physics: Wang-Landau Sampling". In: *American Journal of Physics* 72.10, pp. 1294–1302. doi: 10.1119/1.1707017. URL: <https://doi.org/10.1119/1.1707017>.

Landau, Lev Davidovich and Evgenii Mikhailovich Lifshitz (1980). *Statistical Physics*. 3rd. Vol. 5. Course of Theoretical Physics. Elsevier. doi: 10.1016/C2009-0-24487-4. URL: <https://doi.org/10.1016/C2009-0-24487-4>.

Langton, Christopher G. (1990). "Computation at the Edge of Chaos: Phase Transitions and Emergent Computation". In: *Physica D: Nonlinear Phenomena* 42.1-3, pp. 12–37. doi: 10.1016/0167-2789(90)90064-V. URL: <https://www.sciencedirect.com/science/article/pii/016727899090064V>.

Laughlin, Robert B. and David Pines (2000). "The Theory of Everything". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.1, pp. 28–31. DOI: [10.1073/pnas.97.1.28](https://doi.org/10.1073/pnas.97.1.28). URL: <https://www.pnas.org/doi/10.1073/pnas.97.1.28>.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791). URL: <https://doi.org/10.1109/5.726791>.

LeCun, Yann, Ido Kanter, and Sara A. Solla (1990). "Second Order Properties of Error Surfaces: Learning Time and Generalization". In: *Advances in Neural Information Processing Systems 3*, pp. 918–924. URL: <https://proceedings.neurips.cc/paper/1990/hash/758874998f5bd0c393da094e1967a72b-Abstract.html>.

Leimkuhler, Benedict and Charles Matthews (2013). "Robust and Efficient Configurational Molecular Sampling via Langevin Dynamics". In: *The Journal of Chemical Physics* 138.17, p. 174102. DOI: [10.1063/1.4802990](https://doi.org/10.1063/1.4802990). URL: <https://doi.org/10.1063/1.4802990>.

Levinthal, Cyrus (1969). "How to Fold Graciously". In: *Mössbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois*. Urbana, IL: University of Illinois Press, pp. 22–24. URL: <https://apps.dtic.mil/sti/citations/AD0695202>.

Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw (2010). "Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field". In: *Proteins: Structure, Function, and Bioinformatics* 78.8, pp. 1950–1958. DOI: [10.1002/prot.22711](https://doi.org/10.1002/prot.22711). URL: <https://doi.org/10.1002/prot.22711>.

Linnainmaa, Seppo (1976). "Taylor Expansion of the Accumulated Rounding Error". In: *BIT Numerical Mathematics* 16.2, pp. 146–160. DOI: [10.1007/BF01931367](https://doi.org/10.1007/BF01931367). URL: <https://doi.org/10.1007/BF01931367>.

Liu, Weitang, Yi-Zhuang You, Ying Wai Li, and Jingbo Shang (2023). "Gradient-Based Wang–Landau Algorithm: A Novel Sampler for Output Distribution of Neural Networks over the Input Space". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 22338–22351. URL: <https://proceedings.mlr.press/v202/liu23aw.html>.

Liwo, Adam, Mey Khalili, and Harold A. Scheraga (2005). "Ab Initio Simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains". In: *Proceedings of the National Academy of Sciences of the United*

States of America 102.7, pp. 2362–2367. doi: 10.1073/pnas.0409736102. URL: <https://doi.org/10.1073/pnas.0409736102>.

Liwo, Adam, Jooyoung Lee, Daniel R. Ripoll, Jaroslaw Pillardy, and Harold A. Scheraga (1999). “Protein Structure Prediction by Global Optimization of a Potential Energy Function”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.10, pp. 5482–5485. doi: 10.1073/pnas.96.10.5482. URL: <https://doi.org/10.1073/pnas.96.10.5482>.

Lopez, Michael J. and Shamim S. Mohiuddin (2024). “Biochemistry, Essential Amino Acids”. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing. URL: <https://www.ncbi.nlm.nih.gov/books/NBK557845/>.

Loureiro, Bruno, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová (2021a). “Learning Curves of Generic Features Maps for Realistic Datasets with a Teacher-Student Model”. In: *Advances in Neural Information Processing Systems* 34, pp. 18137–18151. doi: 10.48550/arXiv.2102.08127. URL: <https://proceedings.neurips.cc/paper/2021/hash/9704a4fc48ae88598dcbdcdf57f3fdef-Abstract.html>.

Loureiro, Bruno, Cédric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala (2023). “Fluctuations, Bias, Variance and Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2023.11, p. 114001. doi: 10.1088/1742-5468/ad0221. URL: <https://doi.org/10.1088/1742-5468/ad0221>.

Loureiro, Bruno, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pocco, Florent Krzakala, and Lenka Zdeborová (2021b). “Learning Gaussian Mixtures with Generalized Linear Models: Precise Asymptotics in High-Dimensions”. In: *Advances in Neural Information Processing Systems* 34, pp. 10144–10157. doi: 10.48550/arXiv.2106.03791. URL: <https://proceedings.neurips.cc/paper/2021/hash/543e83748234f7cbab21aa0ade66565f-Abstract.html>.

Machado, Matías R., Exequiel E. Barrera, Florencia Klein, Martín Sónora, Steffano Silva, and Sergio Pantano (2019). “The SIRAH 2.0 Force Field: Altius, Fortius, Citius”. In: *Journal of Chemical Theory and Computation* 15.4, pp. 2719–2733. doi: 10.1021/acs.jctc.9b00006. URL: <https://doi.org/10.1021/acs.jctc.9b00006>.

MacQueen, James B. (1967). “Some Methods for Classification and Analysis of Multivariate Observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Statistics*. Berkeley, CA, USA: University of California Press, pp. 281–297. URL: <https://projecteuclid.org/euclid.bsmsp/1200512992>.

Marrink, Siewert J., H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries (2007). “The MARTINI Force Field: Coarse Grained Model for Biomolec-

ular Simulations". In: *The Journal of Physical Chemistry B* 111.27, pp. 7812–7824. doi: 10.1021/jp071097f. URL: <https://doi.org/10.1021/jp071097f>.

Marrink, Siewert J. and D. Peter Tieleman (2013). "Perspective on the Martini Model". In: *Chemical Society Reviews* 42.16, pp. 6801–6822. doi: 10.1039/C3CS60093A. URL: <https://doi.org/10.1039/C3CS60093A>.

Marsili, Matteo, Iacopo Mastromatteo, and Yasser Roudi (2013). "On Sampling and Modeling Complex Systems". In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.09, P09003. doi: 10.1088/1742-5468/2013/09/P09003. URL: <https://doi.org/10.1088/1742-5468/2013/09/P09003>.

Marsili, Matteo and Yasser Roudi (2022). "Quantifying Relevance in Learning and Inference". In: *Physics Reports* 963, pp. 1–43. doi: 10.1016/j.physrep.2022.03.001. URL: <https://doi.org/10.1016/j.physrep.2022.03.001>.

Marx, Dominik and Jürg Hutter (2009). *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press. doi: 10.1017/CB09780511609633. URL: <https://doi.org/10.1017/CB09780511609633>.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Julia Betts Santarlasci Lapo Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak (2025). *Artificial Intelligence Index Report 2025*. Tech. rep. Stanford Institute for Human-Centered Artificial Intelligence (HAI). URL: <https://hai.stanford.edu/ai-index/2025-ai-index-report>.

Mathur, Apoorva, Rikhia Ghosh, and Ariane Nunes-Alves (2024). "Recent Progress in Modeling and Simulation of Biomolecular Crowding and Condensation Inside Cells". In: *Journal of Chemical Information and Modeling* 64.24, pp. 9063–9081. doi: 10.1021/acs.jcim.4c01520. URL: <https://doi.org/10.1021/acs.jcim.4c01520>.

Matsuda, Hiroyuki (2000). "Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration". In: *Physical Review E* 62.3, pp. 3096–3102. doi: 10.1103/PhysRevE.62.3096. URL: <https://doi.org/10.1103/PhysRevE.62.3096>.

McCammon, James Andrew, Bruce R. Gelin, and Martin Karplus (1977). "Dynamics of Folded Proteins". In: *Nature* 267.5612, pp. 585–590. doi: 10.1038/267585a0. URL: <https://www.nature.com/articles/267585a0>.

McCammon, James Andrew and Stephen C. Harvey (1988). *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press. doi: 10.1017/CB09781139167864. URL: <https://doi.org/10.1017/CB09781139167864>.

- McCulloch, Warren S. and Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5.4, pp. 115–133. DOI: 10.1007/BF02478259. URL: <https://doi.org/10.1007/BF02478259>.
- McGill, William J. (1954). "Multivariate Information Transmission". In: *IRE Transactions on Information Theory* 4.4, pp. 93–111. DOI: 10.1109/TIT.1954.1057469. URL: <https://doi.org/10.1109/TIT.1954.1057469>.
- Mehta, Pankaj, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab (2019). "A High-Bias, Low-Variance Introduction to Machine Learning for Physicists". In: *Physics Reports* 810, pp. 1–124. DOI: 10.1016/j.physrep.2019.03.001. URL: <https://doi.org/10.1016/j.physrep.2019.03.001>.
- Menichetti, Roberto, Marco Giulini, and Raffaello Potestio (2021). "A Journey through Mapping Space: Characterising the Statistical and Metric Properties of Reduced Representations of Macromolecules". In: *The European Physical Journal B* 94.10, p. 204. DOI: 10.1140/epjb/s10051-021-00205-9. URL: <https://doi.org/10.1140/epjb/s10051-021-00205-9>.
- Mézard, Marc and Andrea Montanari (2009). *Information, Physics, and Computation*. Oxford University Press. DOI: 10.1093/acprof:oso/9780198570837.001.0001. URL: <https://doi.org/10.1093/acprof:oso/9780198570837.001.0001>.
- Mézard, Marc, Giorgio Parisi, and Miguel Angel Virasoro (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Lecture Notes in Physics. World Scientific Publishing Company. DOI: 10.1142/0271. URL: <https://doi.org/10.1142/0271>.
- Mignacco, Francesca, Florent Krzakala, Yue M. Lu, Pierfrancesco Urbani, and Lenka Zdeborová (2020). "The Role of Regularization in Classification of High-Dimensional Noisy Gaussian Mixture". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6874–6883. URL: <https://proceedings.mlr.press/v119/mignacco20a.html>.
- Milligan, Glenn W. (1979). "Ultrametric Hierarchical Clustering Algorithms". In: *Psychometrika* 44.3, pp. 343–346. DOI: 10.1007/BF02294699. URL: <https://doi.org/10.1007/BF02294699>.
- Mitsutake, Ayori, Yuji Sugita, and Yuko Okamoto (2001). "Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers". In: *Peptide Science* 60.2, pp. 96–123. DOI: 10.1002/1097-0282(2001)60:2<96::AID-BIP1007>3.0.CO;2-F. URL: [https://doi.org/10.1002/1097-0282\(2001\)60:2%3C96::AID-BIP1007%3E3.0.CO;2-F](https://doi.org/10.1002/1097-0282(2001)60:2%3C96::AID-BIP1007%3E3.0.CO;2-F).

Miyazawa, Sanzo and Robert L. Jernigan (1996). “Residue–Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading”. In: *Journal of Molecular Biology* 256.3, pp. 623–644. doi: 10.1006/jmbi.1996.0114. URL: <https://doi.org/10.1006/jmbi.1996.0114>.

Molchanov, Dmitry, Arsenii Ashukha, and Dmitry Vetrov (2017). “Variational Dropout Sparsifies Deep Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2498–2507. URL: <https://proceedings.mlr.press/v70/molchanov17a.html>.

Monticelli, Luca, Senthil K. Kandasamy, Xavier Periole, Ronald G. Larson, D. Peter Tieleman, and Siewert J. Marrink (2008). “The MARTINI Coarse-Grained Force Field: Extension to Proteins”. In: *Journal of Chemical Theory and Computation* 4.5, pp. 819–834. doi: 10.1021/ct700324x. URL: <https://doi.org/10.1021/ct700324x>.

Mora, Thierry and William Bialek (2011). “Are Biological Systems Poised at Criticality?” In: *Journal of Statistical Physics* 144.2, pp. 268–302. doi: 10.1007/s10955-011-0229-4. URL: <https://doi.org/10.1007/s10955-011-0229-4>.

Morcos, Ari S., David G. T. Barrett, Neil C. Rabinowitz, and Matthew M. Botvinick (2018). “On the Importance of Single Directions for Generalization”. In: *International Conference on Learning Representations (ICLR)*. doi: 10.48550/arXiv.1803.06959. URL: <https://openreview.net/forum?id=rliuQjxCZ>.

Morris, Rhiannon, Katrina A. Black, and Elliott J. Stollar (2022). “Uncovering Protein Function: From Classification to Complexes”. In: *Essays in Biochemistry* 66.3, pp. 255–285. doi: 10.1042/EBC20210044. URL: <https://doi.org/10.1042/EBC20210044>.

Murtagh, Fionn and Pedro Contreras (2012). “Algorithms for Hierarchical Clustering: An Overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1, pp. 86–97. doi: 10.1002/widm.53. URL: <https://doi.org/10.1002/widm.53>.

Nadler, Boaz, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis (2006). “Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems”. In: *Applied and Computational Harmonic Analysis* 21.1, pp. 113–127. doi: 10.1016/j.acha.2005.07.004. URL: <https://doi.org/10.1016/j.acha.2005.07.004>.

Nair, Vinod and Geoffrey E. Hinton (2010). “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814. doi: 10.5555/3104322.3104425. URL: <https://dl.acm.org/doi/10.5555/3104322.3104425>.

Nemenman, Ilya (2011). "Coincidences and Estimation of Entropies of Random Variables with Large Cardinalities". In: *Entropy* 13.12, pp. 2013–2023. doi: 10.3390/e13122013. URL: <https://doi.org/10.3390/e13122013>.

Nemenman, Ilya, Fariel Shafee, and William Bialek (2001). "Entropy and Inference, Revisited". In: *Advances in Neural Information Processing Systems 14 (NeurIPS 2001)*. MIT Press. URL: <https://proceedings.neurips.cc/paper/2001/hash/d46e1fcf4c07ce4a69ee07e4134bcef1-Abstract.html>.

Nightingale, M. Peter and Cyrus J. Umrigar (1998). *Quantum Monte Carlo Methods in Physics and Chemistry*. Vol. 525. NATO Science Series C: Mathematical and Physical Sciences. Springer Science & Business Media. doi: 10.1007/978-94-011-5036-2. URL: <https://doi.org/10.1007/978-94-011-5036-2>.

Noé, Frank and Cecilia Clementi (2017). "Collective Variables for the Study of Long-Time Kinetics from Molecular Trajectories: Theory and Methods". In: *Current Opinion in Structural Biology* 43, pp. 141–147. doi: 10.1016/j.sbi.2017.02.006. URL: <https://doi.org/10.1016/j.sbi.2017.02.006>.

Noé, Frank and Stefan Fischer (2008). "Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules". In: *Current Opinion in Structural Biology* 18.2, pp. 154–162. doi: 10.1016/j.sbi.2008.01.008. URL: <https://doi.org/10.1016/j.sbi.2008.01.008>.

Noé, Frank, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi (2020). "Machine Learning for Molecular Simulation". In: *Annual Review of Physical Chemistry* 71.1, pp. 361–390. doi: 10.1146/annurev-physchem-042018-052331. URL: <https://doi.org/10.1146/annurev-physchem-042018-052331>.

Noid, William George (2013). "Perspective: Coarse-Grained Models for Biomolecular Systems". In: *The Journal of Chemical Physics* 139.9, p. 090901. doi: 10.1063/1.4818908. URL: <https://doi.org/10.1063/1.4818908>.

Noid, William George (2023). "Perspective: Advances, Challenges, and Insight for Predictive Coarse-Grained Models". In: *The Journal of Physical Chemistry B* 127.19, pp. 4174–4207. doi: 10.1021/acs.jpcb.2c08731. URL: <https://doi.org/10.1021/acs.jpcb.2c08731>.

Noid, William George, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen (2008). "The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models". In: *The Journal of Chemical Physics* 128.24, p. 244114. doi: 10.1063/1.2938860. URL: <https://doi.org/10.1063/1.2938860>.

Noid, William George, Ryan J. Szukalo, Katherine M. Kidder, and Maria C. Lesniewski (2024). "Rigorous Progress in Coarse-Graining". In: *Annual Review of*

Physical Chemistry 75.1, pp. 21–45. DOI: 10.1146/annurev-physchem-062123-010821. URL: <https://doi.org/10.1146/annurev-physchem-062123-010821>.

Onuchic, José Nelson, Zaida Luthey-Schulten, and Peter G. Wolynes (1997). “Theory of Protein Folding: The Energy Landscape Perspective”. In: *Annual Review of Physical Chemistry* 48.1, pp. 545–600. DOI: 10.1146/annurev.physchem.48.1.545. URL: <https://doi.org/10.1146/annurev.physchem.48.1.545>.

Onufriev, Alexey, Donald Bashford, and David A. Case (2000). “Modification of the Generalized Born Model Suitable for Macromolecules”. In: *The Journal of Physical Chemistry B* 104.15, pp. 3712–3720. DOI: 10.1021/jp994072s. URL: <https://doi.org/10.1021/jp994072s>.

Opper, Manfred and Ole Winther (2001). “Adaptive and Self-Averaging Thouless-Anderson-Palmer Mean-Field Theory for Probabilistic Modeling”. In: *Physical Review E* 64.5, p. 056131. DOI: 10.1103/PhysRevE.64.056131. URL: <https://doi.org/10.1103/PhysRevE.64.056131>.

Pande, Vijay S., Kyle Beauchamp, and Gregory R. Bowman (2010). “Everything You Wanted to Know about Markov State Models but Were Afraid to Ask”. In: *Methods* 52.1, pp. 99–105. DOI: 10.1016/j.ymeth.2010.06.002. URL: <https://doi.org/10.1016/j.ymeth.2010.06.002>.

Pandya, Maya J., Stefanie Schiffrers, Andrea M. Hounslow, Nicola J. Baxter, and Mike P. Williamson (2018). “Why the Energy Landscape of Barnase Is Hierarchical”. In: *Frontiers in Molecular Biosciences* 5, p. 115. DOI: 10.3389/fmolb.2018.00115. URL: <https://doi.org/10.3389/fmolb.2018.00115>.

Papaleo, Elena, Paolo Mereghetti, Piercarlo Fantucci, Rita Grandori, and Luca De Gioia (2009). “Free-Energy Landscape, Principal Component Analysis, and Structural Clustering to Identify Representative Conformations from Molecular Dynamics Simulations: The Myoglobin Case”. In: *Journal of Molecular Graphics and Modelling* 27.8, pp. 889–899. DOI: 10.1016/j.jmgm.2009.01.006. URL: <https://doi.org/10.1016/j.jmgm.2009.01.006>.

Park, Sanghyun and Klaus Schulten (2004). “Calculating Potentials of Mean Force from Steered Molecular Dynamics Simulations”. In: *The Journal of Chemical Physics* 120.13, pp. 5946–5961. DOI: 10.1063/1.1651473. URL: <https://doi.org/10.1063/1.1651473>.

Parrinello, Michele and Aneesur Rahman (1981). “Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method”. In: *Journal of Applied Physics* 52.12, pp. 7182–7190. DOI: 10.1063/1.328693. URL: <https://doi.org/10.1063/1.328693>.

- Peters, Baron and Bernhardt L. Trout (2006). "Obtaining Reaction Coordinates by Likelihood Maximization". In: *The Journal of Chemical Physics* 125.5, p. 054108. doi: 10.1063/1.2234477. URL: <https://doi.org/10.1063/1.2234477>.
- Poma, Adolfo B., Alejandra Hinostroza Caldas, Luis F. Cofas-Vargas, Michael S. Jones, Andrew L. Ferguson, and Leonardo Medrano Sandonas (2025). "Recent Advances in Machine Learning and Coarse-Grained Potentials for Biomolecular Simulations and Their Applications". In: *Biophysical Journal*. doi: 10.1016/j.bpj.2025.06.019. URL: <https://doi.org/10.1016/j.bpj.2025.06.019>.
- Ponder, Jay W. and David A. Case (2003). "Force Fields for Protein Simulations". In: *Advances in Protein Chemistry* 66, pp. 27–85. doi: 10.1016/S0065-3233(03)66002-X. URL: [https://doi.org/10.1016/S0065-3233\(03\)66002-X](https://doi.org/10.1016/S0065-3233(03)66002-X).
- Pontiggia, Francesco, Giorgio Colombo, Cristian Micheletti, and Henri Orland (2007). "Anharmonicity and Self-Similarity of the Free Energy Landscape of Protein G". In: *Physical Review Letters* 98.4, p. 048102. doi: 10.1103/PhysRevLett.98.048102. URL: <https://doi.org/10.1103/PhysRevLett.98.048102>.
- Pontiggia, Francesco, Andrea Zen, and Cristian Micheletti (2008). "Small- and Large-Scale Conformational Changes of Adenylate Kinase: A Molecular Dynamics Study of the Subdomain Motion and Mechanics". In: *Biophysical Journal* 95.12, pp. 5901–5912. doi: 10.1529/biophysj.108.135467. URL: <https://doi.org/10.1529/biophysj.108.135467>.
- Poppleton, Erik, Michael Matthies, Debesh Mandal, Flavio Romano, Petr Šulc, and Lorenzo Rovigatti (2023). "oxDNA: Coarse-Grained Simulations of Nucleic Acids Made Simple". In: *Journal of Open Source Software* 8.81, p. 4693. doi: 10.21105/joss.04693. URL: <https://doi.org/10.21105/joss.04693>.
- Poulain, Pierre, Florent Calvo, Rodolphe Antoine, Michel Broyer, and Ph. Dugourd (2006). "Performances of Wang-Landau Algorithms for Continuous Systems". In: *Physical Review E* 73.5, p. 056704. doi: 10.1103/PhysRevE.73.056704. URL: <https://doi.org/10.1103/PhysRevE.73.056704>.
- Rahaman, Nasim, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville (2019). "On the Spectral Bias of Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5301–5310. URL: <https://proceedings.mlr.press/v97/rahaman19a.html>.
- Refinetti, Maria, Alessandro Ingrosso, and Sebastian Goldt (2023). "Neural Networks Trained with SGD Learn Distributions of Increasing Complexity". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 28843–28863. URL: <https://proceedings.mlr.press/v202/refinetti23a.html>.

Reith, Dirk, Mathias Pütz, and Florian Müller-Plathe (2003). “Deriving Effective Mesoscale Potentials from Atomistic Simulations”. In: *Journal of Computational Chemistry* 24.13, pp. 1624–1636. DOI: 10.1002/jcc.10307. URL: <https://doi.org/10.1002/jcc.10307>.

Robbins, Herbert and Sutton Monro (1951). “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407. DOI: 10.1214/aoms/1177729586. URL: <https://doi.org/10.1214/aoms/1177729586>.

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi (2024). “Mathematical Discoveries from Program Search with Large Language Models”. In: *Nature* 625.7995, pp. 468–475. DOI: 10.1038/s41586-023-06924-6. URL: <https://doi.org/10.1038/s41586-023-06924-6>.

Rosenblatt, Frank (1958). “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519. URL: <https://doi.org/10.1037/h0042519>.

Roy, Hrittik, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg (2024). “Reparameterization Invariance in Approximate Bayesian Inference”. In: *Advances in Neural Information Processing Systems* 37, pp. 8132–8164. URL: <https://openreview.net/forum?id=204Y0rDHny>.

Rudzinski, Joseph F. and William G. Noid (2011). “Coarse-Graining Entropy, Forces, and Structures”. In: *The Journal of Chemical Physics* 135.21, p. 214101. DOI: 10.1063/1.3663709. URL: <https://doi.org/10.1063/1.3663709>.

Rühle, Victor, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko (2009). “Versatile Object-Oriented Toolkit for Coarse-Graining Applications”. In: *Journal of Chemical Theory and Computation* 5.12, pp. 3211–3223. DOI: 10.1021/ct900369w. URL: <https://doi.org/10.1021/ct900369w>.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). “Learning Representations by Back-Propagating Errors”. In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0. URL: <https://doi.org/10.1038/323533a0>.

Russ, William P. and Rama Ranganathan (2002). “Knowledge-Based Potential Functions in Protein Design”. In: *Current Opinion in Structural Biology* 12.4, pp. 447–452. DOI: 10.1016/S0959-440X(02)00346-9. URL: [https://doi.org/10.1016/S0959-440X\(02\)00346-9](https://doi.org/10.1016/S0959-440X(02)00346-9).

Saad, David and Sara A. Solla (1995). “On-line Learning in Soft Committee Machines”. In: *Physical Review E* 52.4, pp. 4225–4243. DOI: 10.1103/PhysRevE.52.4225. URL: <https://doi.org/10.1103/PhysRevE.52.4225>.

- Sagun, Levent, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou (2017). "Empirical Analysis of the Hessian of Over-Parametrized Neural Networks". In: *arXiv preprint arXiv:1706.04454*. DOI: 10.48550/arXiv.1706.04454. URL: <https://arxiv.org/abs/1706.04454>.
- Sahrmann, Patrick G. and Gregory A. Voth (2025). "On the Emergence of Machine-Learning Methods in Bottom-Up Coarse-Graining". In: *Current Opinion in Structural Biology* 90, p. 102972. DOI: 10.1016/j.sbi.2024.102972. URL: <https://doi.org/10.1016/j.sbi.2024.102972>.
- Samudrala, Ram and John Moult (1998). "An All-Atom Distance-Dependent Conditional Probability Discriminatory Function for Protein Structure Prediction". In: *Journal of Molecular Biology* 275.5, pp. 895–916. DOI: 10.1006/jmbi.1997.1479. URL: <https://doi.org/10.1006/jmbi.1997.1479>.
- Saunders, Marissa G. and Gregory A. Voth (2013). "Coarse-Graining Methods for Computational Biology". In: *Annual Review of Biophysics* 42.1, pp. 73–93. DOI: 10.1146/annurev-biophys-083012-130348. URL: <https://doi.org/10.1146/annurev-biophys-083012-130348>.
- Scheffer, Marten, Steve Carpenter, Jonathan A. Foley, Carl Folke, and Brian Walker (2001). "Catastrophic Shifts in Ecosystems". In: *Nature* 413.6856, pp. 591–596. DOI: 10.1038/35098000. URL: <https://www.nature.com/articles/35098000>.
- Schneirla, Theodore C. (1944). "A unique case of circular milling in ants, considered in relation to trail following and the general problem of orientation". In: *American Museum Novitates* 1253, pp. 1–26. URL: <http://hdl.handle.net/2246/3733>.
- Schwab, David J., Ilya Nemenman, and Pankaj Mehta (2014). "Zipf's Law and Criticality in Multivariate Data without Fine-Tuning". In: *Physical Review Letters* 113.6, p. 068102. DOI: 10.1103/PhysRevLett.113.068102. URL: <https://doi.org/10.1103/PhysRevLett.113.068102>.
- Scott, David W. (1979). "On Optimal and Data-Based Histograms". In: *Biometrika* 66.3, pp. 605–610. DOI: 10.1093/biomet/66.3.605. URL: <https://doi.org/10.1093/biomet/66.3.605>.
- Sengar, Aditya, Thomas E. Ouldridge, Oliver Henrich, Lorenzo Rovigatti, and Petr Šulc (2021). "A Primer on the oxDNA Model of DNA: When to Use It, How to Simulate It and How to Interpret the Results". In: *Frontiers in Molecular Biosciences* 8, p. 693710. DOI: 10.3389/fmolb.2021.693710. URL: <https://doi.org/10.3389/fmolb.2021.693710>.
- Seung, Hyunjun Sebastian, Haim Sompolinsky, and Naftali Tishby (1992). "Statistical Mechanics of Learning from Examples". In: *Physical Review A* 45.8, pp. 6056–6091. DOI: 10.1103/PhysRevA.45.6056. URL: <https://doi.org/10.1103/PhysRevA.45.6056>.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press. doi: 10.1017/CB09781107298019. URL: <https://doi.org/10.1017/CB09781107298019>.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.

Shaw, David E., Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Trevor Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Joseph Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang (2008). "Anton: A Special-Purpose Machine for Molecular Dynamics Simulation". In: *Communications of the ACM* 51.7, pp. 91–97. doi: 10.1145/1364782.1364802. URL: <https://doi.org/10.1145/1364782.1364802>.

Shaw, David E., Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers (2010). "Atomic-Level Characterization of the Structural Dynamics of Proteins". In: *Science* 330.6002, pp. 341–346. doi: 10.1126/science.1187409. URL: <https://www.science.org/doi/10.1126/science.1187409>.

Shell, M. Scott (2008). "The Relative Entropy Is Fundamental to Multiscale and Inverse Thermodynamic Problems". In: *The Journal of Chemical Physics* 129.14, p. 144108. doi: 10.1063/1.2992060. URL: <https://doi.org/10.1063/1.2992060>.

Shell, M. Scott, Pablo G. Debenedetti, and Athanassios Z. Panagiotopoulos (2002). "Generalization of the Wang-Landau Method for Off-Lattice Simulations". In: *Physical Review E* 66.5, p. 056703. doi: 10.1103/PhysRevE.66.056703. URL: <https://doi.org/10.1103/PhysRevE.66.056703>.

Shinzato, Takashi and Yoshiyuki Kabashima (2008). "Perceptron Capacity Revisited: Classification Ability for Correlated Patterns". In: *Journal of Physics A: Mathematical and Theoretical* 41.32, p. 324013. doi: 10.1088/1751-8113/41/32/324013. URL: <https://doi.org/10.1088/1751-8113/41/32/324013>.

Simon, Herbert A. (2012). "The Architecture of Complexity". In: *The Roots of Logistics: A Scientific Perspective*. Springer, pp. 335–361. doi: 10.1007/978-3-642-27922-5_18. URL: https://doi.org/10.1007/978-3-642-27922-5_18.

Sippl, Manfred J. (1990). "Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins". In: *Journal of Molecular Biology* 213.4, pp. 859–883. doi: 10.1016/S0022-2836(05)80269-4. URL: [https://doi.org/10.1016/S0022-2836\(05\)80269-4](https://doi.org/10.1016/S0022-2836(05)80269-4).

- Smith, Paul E. (1999). "The Alanine Dipeptide Free Energy Surface in Solution". In: *The Journal of Chemical Physics* 111.12, pp. 5568–5579. DOI: 10.1063/1.479860. URL: <https://doi.org/10.1063/1.479860>.
- Sokal, Robert R. and Charles D. Michener (1958). "A Statistical Method for Evaluating Systematic Relationships". In: *The University of Kansas Science Bulletin* 38, pp. 1409–1438. URL: <https://www.biodiversitylibrary.org/part/385476>.
- Souza, Paulo C. T., Riccardo Alessandri, Jonathan Barnoud, Sebastian Thallmair, Ignacio Faustino, Fabian Grünewald, Ilias Patmanidis, Haleh Abdizadeh, Bart M. H. Bruininks, and Tsjerk A. Wassenaar (2021). "Martini 3: A General Purpose Force Field for Coarse-Grained Molecular Dynamics". In: *Nature Methods* 18.4, pp. 382–388. DOI: 10.1038/s41592-021-01098-3. URL: <https://doi.org/10.1038/s41592-021-01098-3>.
- Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. DOI: 10.5555/2627435.2670313. URL: <https://jmlr.org/papers/v15/srivastava14a.html>.
- Stansfeld, Phillip J., Joseph E. Goose, Martin Caffrey, Elisabeth P. Carpenter, Joanne L. Parker, Simon Newstead, and Mark S. P. Sansom (2015). "MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes". In: *Structure* 23.7, pp. 1350–1361. DOI: 10.1016/j.str.2015.05.006. URL: <https://doi.org/10.1016/j.str.2015.05.006>.
- Stelter, David and Thomas Keyes (2019). "Simulation of Fluid/Gel Phase Equilibrium in Lipid Vesicles". In: *Soft Matter* 15.40, pp. 8102–8112. DOI: 10.1039/C9SM00854C. URL: <https://doi.org/10.1039/C9SM00854C>.
- Stillinger, Frank H. (1995). "A Topographic View of Supercooled Liquids and Glass Formation". In: *Science* 267.5206, pp. 1935–1939. DOI: 10.1126/science.267.5206.1935. URL: <https://www.science.org/doi/10.1126/science.267.5206.1935>.
- Stillinger, Frank H. and Thomas A. Weber (1984). "Packing Structures and Transitions in Liquids and Solids". In: *Science* 225.4666, pp. 983–989. DOI: 10.1126/science.225.4666.983. URL: <https://www.science.org/doi/10.1126/science.225.4666.983>.
- Stone, John E., David J. Hardy, Ivan S. Ufimtsev, and Klaus Schulten (2010). "GPU-Accelerated Molecular Modeling Coming of Age". In: *Journal of Molecular Graphics and Modelling* 29.2, pp. 116–125. DOI: 10.1016/j.jmgm.2010.06.010. URL: <https://doi.org/10.1016/j.jmgm.2010.06.010>.
- Sugita, Yuji and Yuko Okamoto (1999). "Replica-Exchange Molecular Dynamics Method for Protein Folding". In: *Chemical Physics Letters* 314.1–2, pp. 141–151. DOI:

10.1016/S0009-2614(99)01123-9. URL: [https://doi.org/10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9).

Takada, Shoji (1999). “Gō-ing for the Prediction of Protein Folding Mechanisms”. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.21, pp. 11698–11700. DOI: 10.1073/pnas.96.21.11698. URL: <https://doi.org/10.1073/pnas.96.21.11698>.

Taketomi, Hiroshi, Yuzo Ueda, and Nobuhiro Gō (1975). “Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation: I. The Effect of Specific Amino Acid Sequence Represented by Specific Inter-Unit Interactions”. In: *International Journal of Peptide and Protein Research* 7.6, pp. 445–459. DOI: 10.1111/j.1399-3011.1975.tb02465.x. URL: <https://doi.org/10.1111/j.1399-3011.1975.tb02465.x>.

Tarenzi, Thomas, Giovanni Mattiotti, Marta Rigoli, and Raffaello Potestio (2022). “In Search of a Dynamical Vocabulary: A Pipeline to Construct a Basis of Shared Traits in Large-Scale Motions of Proteins”. In: *Applied Sciences* 12.14, p. 7157. DOI: 10.3390/app12147157. URL: <https://doi.org/10.3390/app12147157>.

Tarenzi, Thomas, Marta Rigoli, and Raffaello Potestio (2021). “Communication Pathways Bridge Local and Global Conformations in an IgG4 Antibody”. In: *Scientific Reports* 11.1, p. 23197. DOI: 10.1038/s41598-021-02323-x. URL: <https://doi.org/10.1038/s41598-021-02323-x>.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (2001). “Estimating the Number of Clusters in a Data Set via the Gap Statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423. DOI: 10.1111/1467-9868.00293. URL: <https://doi.org/10.1111/1467-9868.00293>.

Tirion, Monique M. (1996). “Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis”. In: *Physical Review Letters* 77.9, pp. 1905–1908. DOI: 10.1103/PhysRevLett.77.1905. URL: <https://doi.org/10.1103/PhysRevLett.77.1905>.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). “The Information Bottleneck Method”. In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377. URL: https://www.princeton.edu/~wbialek/our_papers/tishby%2Bal_99.pdf?utm_source=chatgpt.com.

Tokuda, Eric K., Cesar H. Comin, and Luciano da F. Costa (2022). “Revisiting Agglomerative Clustering”. In: *Physica A: Statistical Mechanics and its Applications* 585, p. 126433. DOI: 10.1016/j.physa.2021.126433. URL: <https://doi.org/10.1016/j.physa.2021.126433>.

Torrie, Glenn M. and John P. Valleau (1977). “Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling”. In: *Journal of Com-*

putational Physics 23.2, pp. 187–199. doi: 10.1016/0021-9991(77)90121-8. URL: [https://doi.org/10.1016/0021-9991\(77\)90121-8](https://doi.org/10.1016/0021-9991(77)90121-8).

Tsabar, Michael, Scott B. Lovitch, Ashwini Jambhekar, and Galit Lahav (2021). “Connecting Timescales in Biology: Can Early Dynamical Measurements Predict Long-Term Outcomes?” In: *Trends in Cancer* 7.4, pp. 301–308. doi: 10.1016/j.trecan.2020.12.008. URL: <https://doi.org/10.1016/j.trecan.2020.12.008>.

Tsai, Shan-Ho, Fugao Wang, and David P. Landau (2007). “Critical Endpoint Behavior in an Asymmetric Ising Model: Application of Wang-Landau Sampling to Calculate the Density of States”. In: *Physical Review E* 75.6, p. 061108. doi: 10.1103/PhysRevE.75.061108. URL: <https://doi.org/10.1103/PhysRevE.75.061108>.

Valentim, Alexandra, Julio C. S. Rocha, Shan-Ho Tsai, Ying Wai Li, Markus Eisenbach, Carlos E. Fiore, and David P. Landau (2015). “Exploring Replica-Exchange Wang-Landau Sampling in Higher-Dimensional Parameter Space”. In: *Journal of Physics: Conference Series* 640.1, p. 012006. doi: 10.1088/1742-6596/640/1/012006. URL: <https://doi.org/10.1088/1742-6596/640/1/012006>.

Vapnik, Vladimir N. (1999). “An Overview of Statistical Learning Theory”. In: *IEEE Transactions on Neural Networks* 10.5, pp. 988–999. doi: 10.1109/72.788640. URL: <https://doi.org/10.1109/72.788640>.

Vapnik, Vladimir N. (2013). *The Nature of Statistical Learning Theory*. Springer. doi: 10.1007/978-1-4757-2440-0. URL: <https://doi.org/10.1007/978-1-4757-2440-0>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems* 30, pp. 5998–6008. URL: <https://papers.neurips.cc/paper/7181-attention-is-all-you-need>.

Villa, Elizabeth, Jayati Sengupta, Leonardo G. Trabuco, Jamie LeBarron, William T. Baxter, Tanvir R. Shaikh, Robert A. Grassucci, Poul Nissen, Måns Ehrenberg, and Klaus Schulten (2009). “Ribosome-Induced Changes in Elongation Factor Tu Conformation Control GTP Hydrolysis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.4, pp. 1063–1068. doi: 10.1073/pnas.0811370106. URL: <https://doi.org/10.1073/pnas.0811370106>.

Vogel, Thomas, Ying Wai Li, Thomas Wüst, and David P. Landau (2013). “Generic, Hierarchical Framework for Massively Parallel Wang-Landau Sampling”. In: *Physical Review Letters* 110.21, p. 210603. doi: 10.1103/PhysRevLett.110.210603. URL: <https://doi.org/10.1103/PhysRevLett.110.210603>.

Vogel, Thomas, Ying Wai Li, Thomas Wüst, and David P. Landau (2014). “Scalable Replica-Exchange Framework for Wang-Landau Sampling”. In: *Physical Review E*

90.2, p. 023302. DOI: 10.1103/PhysRevE.90.023302. URL: <https://doi.org/10.1103/PhysRevE.90.023302>.

Volkhardt, Andreas and Helmut Grubmüller (2022). “Estimating Ruggedness of Free-Energy Landscapes of Small Globular Proteins from Principal Component Analysis of Molecular Dynamics Trajectories”. In: *Physical Review E* 105.4, p. 044404. DOI: 10.1103/PhysRevE.105.044404. URL: <https://doi.org/10.1103/PhysRevE.105.044404>.

Voth, Gregory A. (2008). *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC Press. DOI: 10.1201/9781420059564. URL: <https://doi.org/10.1201/9781420059564>.

Wainwright, Martin J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge University Press. DOI: 10.1017/9781108627771. URL: <https://doi.org/10.1017/9781108627771>.

Wales, David J. (2004). *Energy Landscapes: Applications to Clusters, Biomolecules, and Glasses*. Cambridge Molecular Science. Cambridge: Cambridge University Press. DOI: 10.1017/CB09780511721724. URL: <https://doi.org/10.1017/CB09780511721724>.

Wang, Fugao and David P. Landau (2001a). “Determining the Density of States for Classical Statistical Models: A Random Walk Algorithm to Produce a Flat Histogram”. In: *Physical Review E* 64.5, p. 056101. DOI: 10.1103/PhysRevE.64.056101. URL: <https://doi.org/10.1103/PhysRevE.64.056101>.

Wang, Fugao and David P. Landau (2001b). “Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States”. In: *Physical Review Letters* 86.10, pp. 2050–2053. DOI: 10.1103/PhysRevLett.86.2050. URL: <https://doi.org/10.1103/PhysRevLett.86.2050>.

Wang, Jiang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charon, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi (2019). “Machine Learning of Coarse-Grained Molecular Dynamics Force Fields”. In: *ACS Central Science* 5.5, pp. 755–767. DOI: 10.1021/acscentsci.8b00913. URL: <https://doi.org/10.1021/acscentsci.8b00913>.

Warshel, Arieh and Michael Levitt (1976). “Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme”. In: *Journal of Molecular Biology* 103.2, pp. 227–249. DOI: 10.1016/0022-2836(76)90311-9. URL: [https://doi.org/10.1016/0022-2836\(76\)90311-9](https://doi.org/10.1016/0022-2836(76)90311-9).

Watanabe, Satoshi (1960). “Information Theoretical Analysis of Multivariate Correlation”. In: *IBM Journal of Research and Development* 4.1, pp. 66–82. DOI: 10.1147/rd.41.0066. URL: <https://doi.org/10.1147/rd.41.0066>.

Watkin, Timothy L. H., Albrecht Rau, and Michael Biehl (1993). "The Statistical Mechanics of Learning a Rule". In: *Reviews of Modern Physics* 65.2, pp. 499–556. DOI: 10.1103/RevModPhys.65.499. URL: <https://doi.org/10.1103/RevModPhys.65.499>.

Watts, Duncan J. and Steven H. Strogatz (1998). "Collective Dynamics of 'Small-World' Networks". In: *Nature* 393.6684, pp. 440–442. DOI: 10.1038/30918. URL: <https://www.nature.com/articles/30918>.

Weiner, Scott J., Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Giuliano Alagona, Salvatore Profeta, and Paul Weiner (1984). "A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins". In: *Journal of the American Chemical Society* 106.3, pp. 765–784. DOI: 10.1021/ja00315a051. URL: <https://doi.org/10.1021/ja00315a051>.

Werbos, Paul J. (1982). "Applications of Advances in Nonlinear Sensitivity Analysis". In: *System Modeling and Optimization*. Vol. 38. Lecture Notes in Control and Information Sciences. Berlin, Heidelberg: Springer, pp. 762–770. DOI: 10.1007/BFb0006203. URL: <https://doi.org/10.1007/BFb0006203>.

Wilson, David Sloan (1975). "A Theory of Group Selection". In: *Proceedings of the National Academy of Sciences of the United States of America* 72.1, pp. 143–146. DOI: 10.1073/pnas.72.1.143. URL: <https://doi.org/10.1073/pnas.72.1.143>.

Worschech, Roman and Bernd Rosenow (2024). "Asymptotic Generalization Errors in the Online Learning of Random Feature Models". In: *Physical Review Research* 6.2, p. L022049. DOI: 10.1103/PhysRevResearch.6.L022049. URL: <https://doi.org/10.1103/PhysRevResearch.6.L022049>.

Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). "Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms". In: *CoRR* abs/1708.07747. DOI: 10.48550/arXiv.1708.07747. URL: <https://arxiv.org/abs/1708.07747>.

Yan, Qiliang and Juan J. de Pablo (2003). "Fast Calculation of the Density of States of a Fluid by Monte Carlo Simulations". In: *Physical Review Letters* 90.3, p. 035701. DOI: 10.1103/PhysRevLett.90.035701. URL: <https://doi.org/10.1103/PhysRevLett.90.035701>.

Yin, Junqi and David P. Landau (2012). "Massively Parallel Wang–Landau Sampling on Multiple GPUs". In: *Computer Physics Communications* 183.8, pp. 1568–1573. DOI: 10.1016/j.cpc.2012.02.023. URL: <https://doi.org/10.1016/j.cpc.2012.02.023>.

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson (2014). "How Transferable Are Features in Deep Neural Networks?" In: *Advances in Neural Information*

Processing Systems 27, pp. 3320–3328. URL: <https://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks>.

Zdeborová, Lenka (2020). “Understanding Deep Learning Is Also a Job for Physicists”. In: *Nature Physics* 16.6, pp. 602–604. DOI: 10.1038/s41567-020-0929-2. URL: <https://doi.org/10.1038/s41567-020-0929-2>.

Zhou, Chenggang and R. N. Bhatt (2005). “Understanding and Improving the Wang-Landau Algorithm”. In: *Physical Review E* 72.2, p. 025701. DOI: 10.1103/PhysRevE.72.025701. URL: <https://doi.org/10.1103/PhysRevE.72.025701>.

Zhou, Chenggang, Thomas C. Schulthess, Stefan Torbrügge, and David P. Landau (2006). “Wang-Landau Algorithm for Continuous Models and Joint Density of States”. In: *Physical Review Letters* 96.12, p. 120201. DOI: 10.1103/PhysRevLett.96.120201. URL: <https://doi.org/10.1103/PhysRevLett.96.120201>.

Zhou, Chenggang and Jia Su (2008). “Optimal Modification Factor and Convergence of the Wang-Landau Algorithm”. In: *Physical Review E* 78.4, p. 046705. DOI: 10.1103/PhysRevE.78.046705. URL: <https://doi.org/10.1103/PhysRevE.78.046705>.

Zipf, George Kingsley (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press. DOI: 10.4159/harvard.9780674434929. URL: <https://doi.org/10.4159/harvard.9780674434929>.