INTERFACE FOCUS

royalsocietypublishing.org/journal/rsfs

Research

Cite this article: Giulini M, Potestio R. 2019 A deep learning approach to the structural analysis of proteins. *Interface Focus* **9**: 20190003.
http://dx.doi.org/10.1098/rsfs.2019.0003

Accepted: 13 March 2019

One contribution of 15 to a theme issue 'Multi-resolution simulations of intracellular processes'.

Subject Areas:
biophysics, bioinformatics, computational biology

Keywords:
deep neural networks, protein structure, elastic network models

Author for correspondence:
Raffaello Potestio
e-mail: raffaello.potestio@unitn.it

THE ROYAL SOCIETY PUBLISHING

A deep learning approach to the structural analysis of proteins

Marco Giulini[1,2] and Raffaello Potestio[1,2]

[1]Physics Department, University of Trento, via Sommarive 14, 38123, Trento, Italy
[2]INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, 38123 Trento, Italy

RP, 0000-0001-6408-9380

Deep learning (DL) algorithms hold great promise for applications in the field of computational biophysics. In fact, the vast amount of available molecular structures, as well as their notable complexity, constitutes an ideal context in which DL-based approaches can be profitably employed. To express the full potential of these techniques, though, it is a prerequisite to express the information contained in a molecule's atomic positions and distances in a set of input quantities that the network can process. Many of the molecular descriptors devised so far are effective and manageable for relatively small structures, but become complex and cumbersome for larger ones. Furthermore, most of them are defined locally, a feature that could represent a limit for those applications where global properties are of interest. Here, we build a DL architecture capable of predicting non-trivial and intrinsically global quantities, that is, the eigenvalues of a protein's lowest-energy fluctuation modes. This application represents a first, relatively simple test bed for the development of a neural network approach to the quantitative analysis of protein structures, and demonstrates unexpected use in the identification of mechanically relevant regions of the molecule.

1. Introduction

Proteins are the most versatile biological molecules, as they cover roles ranging from 'mere' structural support (e.g. in the cytoskeleton) to active cargo transport, passing through enzymatic chemistry, protein folding chaperoning, communication, photochemical sensing, etc. The impressive variety of activities, sizes, shapes and functions proteins show is largely due to the LEGO-like capacity of the polypeptide chain, as well as to the polymorphic chemistry entailed in the 20 amino acids they are made of. According to the well-established central dogma of biology, the amino acid sequence of the protein dictates its three-dimensional structure, which in turn determines and enables the molecule's function. It should thus come as no surprise that protein structures have been thoroughly investigated at all levels, from the fundamental, experimental determination of the arrangement of their atoms in space (e.g. by means of X-ray crystallography or nuclear magnetic resonance) to computer-aided analyses aimed at understanding the interplay between sequence, structure and function. These latter studies are carried out through *in silico* representations of the molecules whose resolution ranges from atomistic—as is typically the case in molecular dynamics (MD) [1,2]—to simplified, *coarse-grained* models [3–6], where several atoms are lumped together in sites interacting via effective potentials. Furthermore, the field of protein bioinformatics has boomed in the past few decades, where the wealth of available sequences and structures has been exploited to investigate structure prediction, protein–protein interaction, docking, protein-related genomics, etc. (see [7,8] for recent, comprehensive reviews).

The availability of a large number of instances of the protein space (be that sequence or structure) and the necessity to perform dataset-wide analyses and screening of their properties naturally leads one to wonder whether one could

take advantage of the recent progresses achieved by machine learning approaches, in particular, deep learning (DL). The latter is a subset of the wide class of machine learning computational methods, and has been successfully applied to a fairly wide spectrum of areas of science [9,10], ranging from neuroscience [11] to image and speech recognition [12,13]. In the field of computational chemistry much effort has been devoted to the identification of the variables that are able to provide a comprehensive description of a chemical compound (*molecular descriptors*). These features are usually designed in order to be applied to elements of the chemical compound space (CCS), the theoretical combinatorial set of all possible compounds that could be isolated and constructed from all combinations and configurations of atoms [14–16]. Several examples of descriptors are present in the literature [17–23]: they proved to be extremely useful in the development of predictive models about a huge variety of molecular properties. Nevertheless, the size of the CCS is limited from above by the Lipinski rules [24,25], that set the maximum molecular weight to 500 atomic mass units. It thus appears evident that the vast majority of structures studied in biophysics, such as proteins, nucleic acids and polysaccharides, falls well beyond this value. As an example, the structures conserved in the Protein Data Bank (PDB) have a molecular weight ranging from few hundreds to hundred thousand daltons. One of the biggest issues in the application of DL-based approaches to biophysical problems thus consists in defining a flexible and robust method to properly encode the huge amount of information contained in these structures (*feature extraction*).

Nonetheless, DL algorithms are enjoying increasing popularity in the context of biological and condensed matter physics as well. For example, Geiger & Dellago [26] built DL architectures capable of identifying local phases in liquids. More recently, Feinberg *et al.* [27] developed PotentialNet, a DL-based model for predicting molecular properties. At the same time, Wehmeyer & Noé [28] have implemented a complex DL algorithm (time-lagged autoencoder) able to perform efficient dimensionality reduction on molecular dynamics trajectories. For what concerns the field of protein folding, Wang *et al.* [29,30] developed deep convolutional architectures in order to predict secondary structure and residue–residue contacts from sequence. Notably, promising works by Lemke & Peter [31] and Zhang *et al.* [32] extended the use of such algorithms to the field of coarse-grained models.

In spite of the recent encouraging attempts, a straightforward approach to a DL-based structural analysis protocol for the study of large macromolecules is lacking. In the present work, we aim at moving a step forward in this direction through the construction of a DL model to analyse protein structures. We set ourselves a relatively simple goal, that is, to predict the 10 lowest eigenvalues of an exactly solvable coarse-grained model of a protein's fluctuations. These quantities, in fact, are related to those deformations which involve a large number of atoms moving in a concerted manner; the low eigenvalues associated with such collective movements are indicative of a small amount of energy required to excite them, as well as of the long self-correlation time that it takes to the molecule to relax these deformations back to the equilibrium structure [33,34]. The low-energy eigenvalues are thus simple and representative proxies for the most collective and global properties of the model under examination.

The first aim of this work consists in identifying the procedure of feature extraction that is most suitable to our task. Second, we show that the application of a simple, standard and computationally not expensive DL architecture to the selected features gives satisfactory results, suggesting that more complex tasks will be attainable with similar, more refined networks. It is worth pointing out that although the development of a DL-based predictive model leads to a significant computational gain with respect to the exact algorithmic procedure, this is not the purpose of this work: here, we focus on demonstrating the viability of a DL-based approach to a specific class of problems in computational biophysics.

2. Material and methods

In this section, we first summarize a few relevant concepts about deep neural networks (DNNs), and specifically on convolutional neural networks (CNNs). Subsequently, we provide a brief overview of the protein models of interest for our work, that is, elastic network models (ENMs).

The raw data employed in the present work, including PDB files, protein structure datasets, CNN training protocols, trained networks and related material are publicly available on the ERC VARIAMOLS project website http://variamols.physics.unitn.eu in the research output section.

2.1. Convolutional neural network model

Born in the 1950s as theoretical, simplified models of neural structure and activity, neural networks are becoming an increasingly pervasive instrument for the most diverse types of computation. In particular, the tasks in which DNNs excel are those that can be reduced to classification, pattern recognition, feature extraction and, more recently, even a rudimentary (yet impressive) creative process. DNNs can be understood as a very complex form of *fitting procedure*, in that the parameters of the network are set through a process of training over a large dataset of items for which the outcome value is known; a prototypical example is that of a network endowed with the task of distinguishing images of dogs from those of cats, which is 'trained' by proposing to it several images of the two types and changing the parameters so that the outcome label corresponds to the correct one. The following step is the validation of the network's effectiveness onto a complementary dataset of input instances that have not been employed in the training. The predictive power of DNNs is largely due to the nonlinear character of the functions employed to connect one 'neuron' to the following. This characteristic makes them substantially more flexible and versatile than multi-dimensional linear regression models, albeit also more obscure to comprehend in their functioning.

Deep feedforward networks (also called multilayer perceptrons (MLPs) or artificial neural networks (ANNs)) are the most known class of machine learning algorithms [10]. Given some input values $\boldsymbol{x}$ and an output label y (categorical or numerical), in MLPs we assume the existence of a stochastic function F of x such that $y = F(\boldsymbol{x})$. A mapping $y = f(\boldsymbol{x}; \boldsymbol{W})$ is defined and the algorithm attempts to learn the values of parameters $\boldsymbol{W}$ that give the best approximation of F. This function $f(\boldsymbol{x}; \boldsymbol{W})$ is the composition of n different functions (usually called *layers*), where n is the depth of the MLP:

$$f(\boldsymbol{x}) = f_n(f_{n-1}(\cdots f_2(f_1(\boldsymbol{x})))). \tag{2.1}$$

The function f_1, which directly acts on the input data, is called the *input layer*, while f_n is the *output layer*. $f_2 \cdots f_{n-1}$ and, more generally, all the intermediate layers are called *hidden* because their scope is to translate the results coming from the first layer into an input that can be processed by the output layer.

Equation (2.1) shows that a layer can be thought of as a function that takes a vector as input and gives a different vector as output. One can also imagine a layer as a set of vector-to-scalar

functions (neurons) that act in parallel [10]. Neurons are the building blocks of an MLP. These entities loosely resemble their analogous biological counterpart: each unit receives a certain amount of input signals from other units, adds a custom bias term, performs a weighted sum and applies a nonlinear transformation, or *activation function*, in order to produce an output signal. This transformation has to be nonlinear: in fact, a neural network with only linear activations in the hidden layers would be equivalent to a linear regression model [10].

Among the several neural network architectures that have been developed throughout the years, a particular class is that of convolutional neural networks (CNNs). CNNs proved to be extremely powerful if applied to processes like image and video recognition and natural language processing.

Mathematically, the bidimensional discrete convolution between two functions F and G is given by the following expression:

$$(F \circledast G)(i, j) = \sum_{m,n=-\infty}^{\infty} F(m, n)G(i - m, j - n), \tag{2.2}$$

where F is referred to as the input function (a bidimensional grid-like object) while G is called *kernel function*. G is much smaller than F.

In the vast majority of CNNs [10,13,35], a convolutional layer does not contain only the convolution operation: in fact, it is followed by an activation layer and, usually, by a pooling layer. The activation layer transforms the feature map through the application of a nonlinear function [10]. Pooling layers downscale the input data: given the output of the activation layer at a certain location, a pooling operation performs a summary statistic (average, maximum [36]) on its neighbours that replaces the original value.

Common CNNs consist of a sequence of convolutional layers followed by a number of *fully connected* (*dense*) layers placed before the output. This is the network architecture of choice for this work, as detailed later on.

2.2. Elastic network models

Classical MD [1,2], by which Newton's equations of motion are numerically integrated, is the most effective and widespread method used to investigate *in silico* the equilibrium properties and the dynamics of a (biological) molecule. Despite the recent dramatic gains in computational efficiency [37–40], many biological phenomena cannot be investigated with atomically detailed models: this is a particularly limiting problem if the system size exceeds a few tens of millions of atoms or if the relevant biological processes occur over long time scales (typically larger than hundreds of microseconds). Furthermore, it is important to underline that highly detailed atomistic MD simulations generate an enormous amount of data, which are often difficult to store and post-process and, sometimes, simply not needed.

MD simulations rely on sophisticated semi-empirical potentials that depend on a large number of parameters and reference properties; however, in a seminal paper Tirion [33] showed that, in several cases, it is possible to replace the atomistic potential with a much simpler, single-parameter harmonic spring:

$$E_P = \sum_{(i,j)} E(\boldsymbol{r}_i, \boldsymbol{r}_j) = \sum_{(i,j)} \frac{C}{2}(|\boldsymbol{r}_{i,j}| - |\boldsymbol{r}^0_{i,j}|)^2 \tag{2.3}$$

where the parentheses in the summation (i, j) indicate that the sum is restricted to those atom pairs whose distance $|\boldsymbol{r}_{i,j}| = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is lower than a cut-off radius.

This functional form of the potential is extremely simplistic, as 3-body terms are not even taken into consideration. Nevertheless, it can capture the collective, low energy vibrations of proteins. The slowest modes of vibration involve several atoms and interatomic interactions, whose sum approaches a universal form governed by the central limit theorem. For slow, collective modes, the details of the form of the pair potentials can be neglected [33], and if one is only interested in analysing these modes (which usually dictate the function-oriented dynamics of the molecule) a single-parameter harmonic description can provide accurate predictions.

The potential energy in equation (2.3) gives rise to the following Hessian matrix:

$$\begin{aligned} M_{ij,\mu\nu} &= \frac{\partial^2 V}{\partial x_{i,\mu} \partial x_{j,\nu}} \\ &= -k_{\mathrm{ANM}} \frac{(x_{i,\mu} - x_{j,\mu})(x_{i,\nu} - x_{j,\nu})}{|\boldsymbol{r}^0_{(i,j)}|^2}, \end{aligned} \tag{2.4}$$

where $\boldsymbol{x}_i = \boldsymbol{r}_i - \boldsymbol{r}^0_i$ and μ and ν are Cartesian components. Models described solely by the Hessian matrix in equation (2.4) are called anisotropic elastic network models, or ANMs. The advantage of a quadratic approximation to equation (2.3) is that the normal modes of vibration can be straightforwardly obtained through the inversion of the Hessian.

As anticipated, ENMs can be employed in contexts other than the analysis of vibrational spectra. In fact, it is possible to associate the harmonic force field of these models with simplified representations of the structure, that is, coarse-grained models. Coarse-graining can be defined as the process of reducing the accuracy and resolution of the representation of a system, describing it in terms of fewer collective degrees of freedom and effective interactions. The former are usually defined lumping together a relatively large number of atoms (2–3 to tens) into a single bead; the latter, on the other hand, are parametrized making use of one of the many available strategies [3–6], which in general aim at reproducing the multi-body potential of mean force of the system. Coarse-grained ENMs are typically constructed retaining only the C_α atom of the backbone, and placing a harmonic spring between pairs of atoms whose distance in the native conformation lies within a given interaction cut-off. More refined models employ also the C_β carbon atom—or an equivalent one—which explicitly accounts for the amino acid side chain. It is important to underline that the spring potentials employed in coarse-grained ENMs are a proxy for a thermal average of true all-atom interactions over all conformations compatible with a given coarse-grained configuration; hence, they consist of free energies rather than potential energies, as is usually the case in the context of coarse-graining.

Studies making use of all-atom or coarse-grained ENMs proved to be particularly effective in the modelling and prediction of low energy conformational fluctuations, corresponding to the most collective normal modes. These results are often in agreement with the ones produced using all-atom MD simulations with a standard semi-empirical force field. Among the most notable structures they have been applied to we point out RNA Polymerase II [41], virus capsids [42], transmembrane channels [43] and the whole ribosome [44].

The β-Gaussian model (β-GM [34]) is a particular flavour of coarse-grained ENM in which the description of protein fluctuations is improved through the introduction of effective C_β centroids (with the exception of glycine residues, whose side chain is made up by a single hydrogen atom). The β-GM model is defined on a coarse-grained protein structure, thus providing a simplified description of the system's fluctuations in a local free energy minimum, centred on a reference structure typically chosen to be the native, crystallographic conformation. The introduction of C_β atoms in the model results in a Hamiltonian that is considerably more complex than the one relative to a chain of C_α units, whose general form is given by

$$\begin{aligned} \mathcal{H} = &\frac{1}{2}\sum_{i,j} \boldsymbol{x}_i^{C_\alpha} \boldsymbol{M}_{i,j}^{C_\alpha - C_\alpha} \boldsymbol{x}_j^{C_\alpha} + \sum_{i,j} \boldsymbol{x}_i^{C_\alpha} \boldsymbol{M}_{i,j}^{C_\alpha - C_\beta} \boldsymbol{x}_j^{C_\beta} \\ &+ \frac{1}{2}\sum_{i,j} \boldsymbol{x}_i^{C_\beta} \boldsymbol{M}_{i,j}^{C_\beta - C_\beta} \boldsymbol{x}_j^{C_\beta}, \end{aligned} \tag{2.5}$$

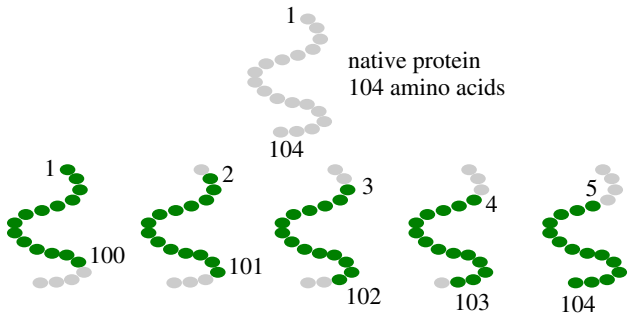

Figure 1. Schematic illustration of the procedure followed to construct the 100-amino acids long decoys. Proteins whose sequence is longer than 100 residues are cut in 100-residues long subsequences sliding a window of this length along the main chain. A protein of length $100 + N$ amino acids produces $N + 1$ decoys. (Online version in colour.)

where M is an interaction matrix. The first term of the Hamiltonian represents the interactions between C_α atoms, be they bonded links between consecutive C_α's along the peptide chain or simply those belonging to close-by residues in contact in the native conformation. The second term accounts for interactions between C_α's and C_β's; lastly, the third term includes interactions existing solely among C_β's.

In the β-GM framework, the positions of C_β's are assigned using a simplified version of the Park–Levitt procedure [45], in which the C_β centroids are placed in the plane specified by the local C_α trace. This assumption allows one to compute the C_β coordinates of atom i using only the positions of C_α's $i-1$, i and $i+1$, giving rise to a Hamiltonian that is quadratic in the C_α deviations but with a different coupling matrix. This model has the very same computational cost of less accurate, C_α-only anisotropic models, but it is able to capture in a more accurate way the low-energy macromolecular fluctuations. In this way, the deviations of C_β atoms of all amino acids (excluding glycine and the terminal residues) are parametrized using the C_α Cartesian coordinates, leading to a new Hamiltonian of the form

$$\tilde{\mathcal{H}} = \frac{1}{2}\sum_{i,j} \boldsymbol{x}_i^{C_\alpha} \tilde{\boldsymbol{M}}_{i,j}^{C_\alpha - C_\alpha} \boldsymbol{x}_j^{C_\alpha}. \tag{2.6}$$

In the present work, we have been consistent with the model as described in the original paper [34], and used the same parameters present therein. In particular, the cut-off radius R_c has been set to 7.5 Å.

2.3. Construction of the protein dataset

In the previous section, we described the exactly solvable algorithmic procedure through which one can compute eigenvalues and eigenvectors associated with the local fluctuation dynamics of the β-GM coarse-grained protein model. As anticipated in the introduction, the scope of our work consists in building a DL architecture (CNN) capable of predicting the lowest 10 of these eigenvalues. In order to do so we have to first train and subsequently validate this CNN approach. We constructed two separate groups of protein structures, downloaded from the PDB, to be used as training and evaluation sets, respectively. The evaluation set contains protein structures with a single chain and 100 C_α atoms; for the training set, we considered chains with a length between 101 and 110 monomers that have been processed to construct $N+1$ *decoys* for each protein of length $100+N$. In this specific context, by decoy, we indicate protein-like chains or subchains that preserve the vast majority of typical structural properties of real, 'full' proteins [46]. Figure 1 illustrates the procedure followed to produce such decoys.

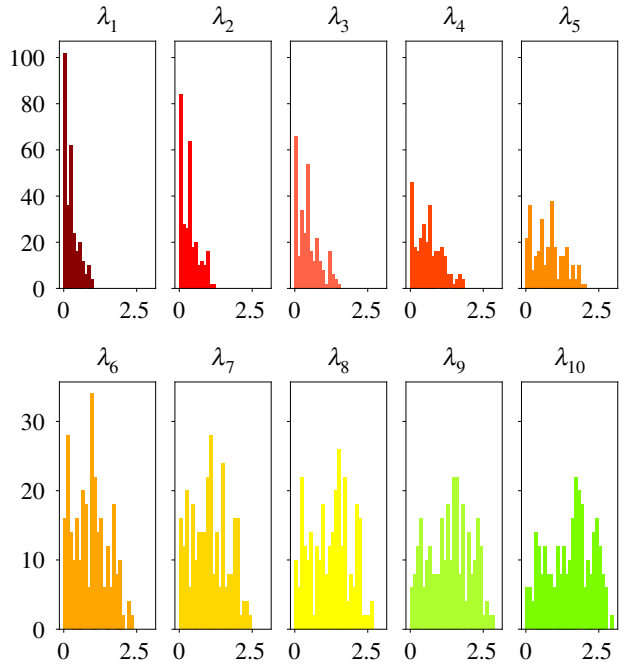

Figure 2. Distribution of the lowest 10 eigenvalues λ of the spectrum of all proteins in the evaluation set, computed by means of the β-GM. (Online version in colour.)

Through this simple process, we obtain an evaluation set of 146 real proteins with 100 amino acids and a training set of 10728 decoys of the same length. It is important to note that the β-GM spectrum is invariant with respect to the *orientation* of the sequence, namely we can easily double both datasets including the reversed structures.

Dealing with proteins and biologically relevant decoys we encounter a wide variety of structures. They are extremely heterogeneous in terms of radius of gyration and their spectra show high variability. Figure 2 shows the distribution of the 10 lowest eigenvalues in the validation set, where λ_i represents the i-th smallest non-zero eigenvalue, $i = 1, 2, \ldots 10$. The choice of considering only 10 eigenvalues is customary and in line with the analyses carried out in the literature [47–50].

In order to make a quantitative comparison between samples in the available datasets, figure 3 shows a histogram of the globularity expressed in terms of radius of gyration

$$R_g = \sqrt{\frac{1}{N}\sum_{k=1}^{N} (\boldsymbol{r}_k - \boldsymbol{r}_{\text{mean}})^2}. \tag{2.7}$$

In the training set, there are several structures that are highly non-globular, but the vast majority has a gyration radius comparable to the values present in the evaluation set.

2.4. Molecular descriptor

A crucial step in the construction of a DL-based pipeline to analyse and process a given molecular structure is the identification of an appropriate *molecular descriptor*. From equation (2.6), we can see that, within the β-GM framework, the Hamiltonian of the system depends only on the positions of the C_α atoms. Hence, our molecular descriptor will take as input only the Cartesian coordinates of these atoms. However, for CNN applications, we cannot simply characterize the biomolecule in terms of Cartesian coordinates, since these are not invariant with respect to rotations and translations of the system, an important requirement a molecular descriptor has to fulfil.

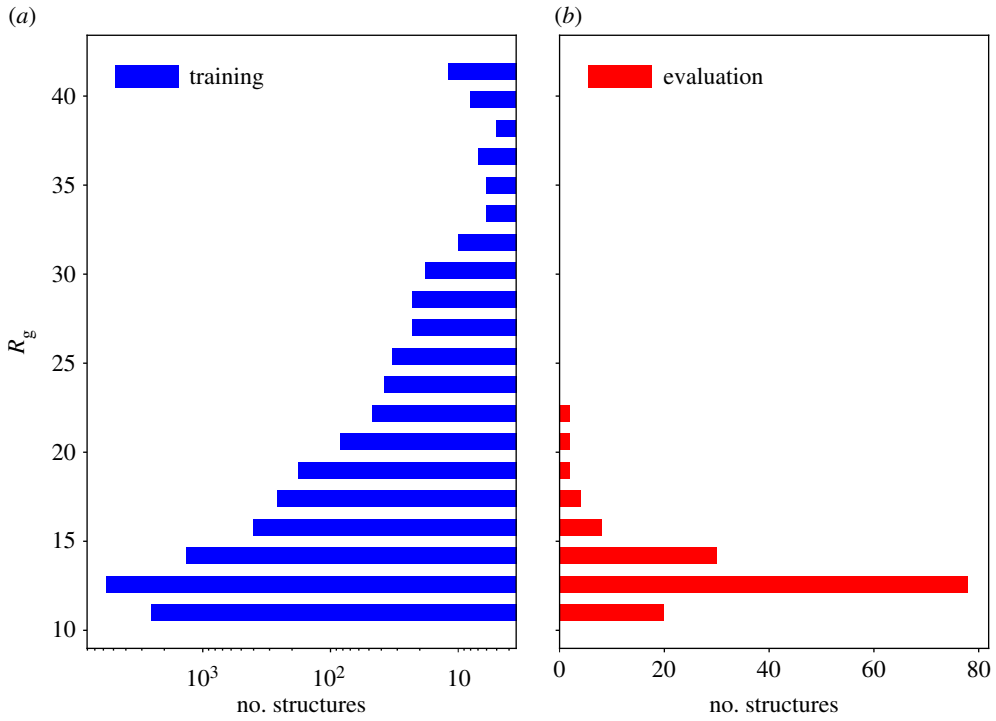

Figure 3. Distribution of the radius of gyration over training set (*a*) and evaluation set (*b*). (Online version in colour.)

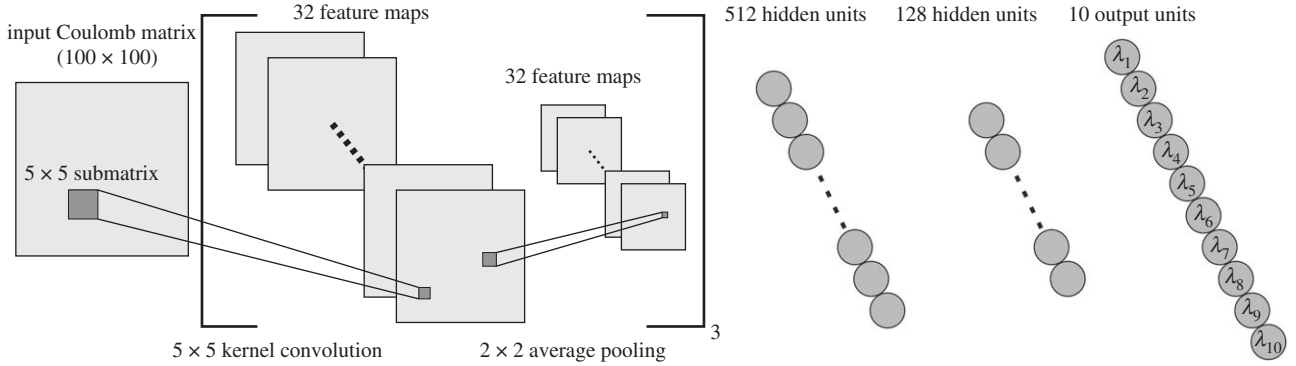

Figure 4. The CNN architecture employed in the present work. A convolutional layer is drawn between squared brackets to emphasize that it is iteratively repeated for three times. For the sake of clarity, connections between neurons in the fully connected layers are not represented. (Online version in colour.)

A prominent example of molecular descriptor is given by Behler and Parrinello's *symmetry functions* [51,52]. These functions describe the local environment of each atom in a molecular system, while satisfying the invariance requirement. Among the parameters that are defined in the calculation of these quantities, the most relevant one from a conceptual point of view is the cut-off radius: interatomic distances larger than this value yield zero contribution to these descriptors. Symmetry functions have been mainly employed in order to provide accurate potential energy surfaces [51,52] and to detect local atomic arrangements in liquids [26]. Although these descriptors proved to be extremely successful, our learning task concerns the prediction of a property that is (at least in general) intrinsically global, hence we need a function that is able to encode all the interactions between the atoms that constitute the system. Therefore, we decided to characterize the proteins under examination in terms of the Coulomb matrix, a very general and global molecular descriptor that is rotation-translation invariant. This is defined as

$$C_{IJ} = \begin{cases} 0 & \text{if } I = J, \\ \frac{1}{|R_I - R_J|} & \text{otherwise,} \end{cases} \tag{2.8}$$

where I and J are atomic indices.

2.5. Architecture

In the previous sections, we defined all the elements of our learning problem, namely the chosen molecular descriptor, the desired output and the algorithmic procedure used to produce it. We now illustrate the architecture of the network employed.

The motivations behind the choice of a CNN architecture to address the problem at hand are essentially three. First, *parameter sharing* allows one to keep the total number of parameters to be *learned* relatively low. If we used an ANN, which has *tied* weights, we would have obtained a much higher total number of learnable parameters. Second, CNNs are particularly suited to deal with grid-like input data, such as Coulomb matrices. Third, no data preprocessing is required.

Here, we used a CNN composed of three convolutional layers and two fully connected layers. Each convolutional layer is made by a convolution operation followed by an average pooling layer. While the latter acts on regions of amplitude 2×2, the former is realized with the use of 32 kernel functions, each of which is a 5×5 matrix whose elements represent the learnable parameters (weights). The dense layers consist of 512 and 128 units, respectively. There are 10 output units, each of which corresponds to one non-zero eigenvalue of the β-GM spectrum. The network structure is sketched in figure 4. Three *dropout* [53]

layers have been included in the network before, between, and after the fully connected layers. Dropout is a *regularization* technique that drops a certain ratio (25% in our case) of the input units of a layer at each step of the training process. This technique significantly prevents the risk of overfitting the training set [54].

The network is developed using Keras [55] with Tensorflow [56] backend. The optimizer is adagrad [57] with learning rate = 0.008. The batch size is 400 and the number of epochs is bound to 100. For what concerns the loss function, we decided to employ the mean absolute percentage error (MAPE):

$$\text{MAPE} = 100 \times \frac{1}{N}\sum_{j=1}^{N}\frac{1}{10}\sum_{i=1}^{10}\frac{|\lambda_i^j - \hat{\lambda}_i^j|}{|\hat{\lambda}_i^j|}, \tag{2.9}$$

where N is the batch size while $\hat{\lambda}_i^j$ and λ_i^j represent true and predicted eigenvalues, respectively. Recent work by de Myttenaere *et al*. [58] proved generic consistency results for this loss function.

In order to quantitatively assess the effectiveness of the network, we analysed the CNN-predicted eigenvalues through a cross-validation procedure. This is a common strategy to evaluate the performances of a learning algorithm and its ability to generalize to an unknown and independent dataset. The idea behind this technique is the repetition of training and testing processes on different subsets of the full training set. k-fold cross-validation is the most known example of this procedure: the full training set is split into k different *folds*; for each of these subsets the algorithm is trained over the other $k-1$ folds and is tested against the unknown samples present in the k-th fold. In this work, we have made use of the deep analysis protocol (DAP) for cross-validation. This protocol has been extensively employed in many machine learning challenges applied to biological data [59,60], inducing an effective massive replication of data. In this work, we performed a 10×5 cross-validation, namely a fivefold cross-validation performed 10 times, with 10 different random seeds for the network; the latter seeds are the same that have been used during the process of training on the full dataset. This iterative partitioning of the protein decoys dataset into training and validation subsets thus allows us to assess the network's predictive power onto a minimally overlapping group of structures, while at the same time taking full advantage of the number of available input elements for the parametrization procedure.

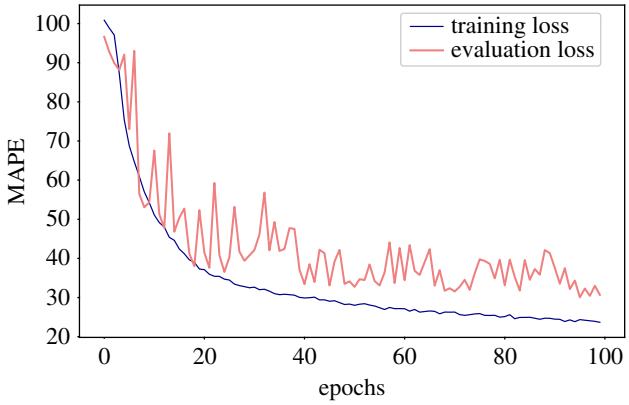

Figure 5. Example of the behaviour of losses against the number of epochs. Both training loss and evaluation loss refer to the MAPE (equation (2.9)), computed on the samples in the training and evaluation set, respectively. (Online version in colour.)

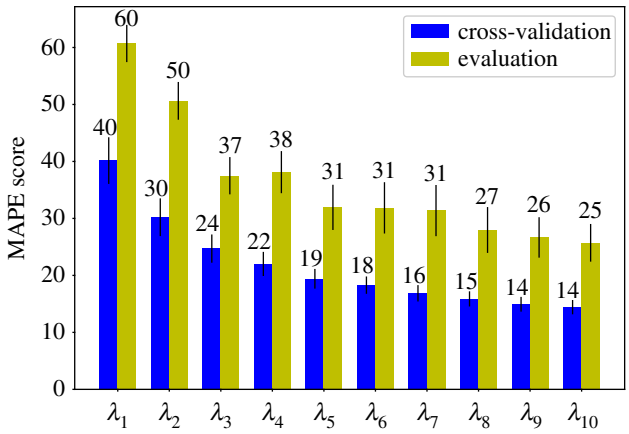

Figure 6. MAPE scores as obtained for simple evaluation (green bars) and cross-validation (blue bars), broken down to each eigenvalue separately. The specific MAPE value of each bar is indicated above the latter. (Online version in colour.)

3. Results and discussion

Before discussing the results we deem it useful to highlight a few crucial aspects of the purpose of our work. In essence, the problem we tackle here can be seen as a spectral inversion by means of a CNN. In the literature, there are previous examples [61,62] of machine learning-based approaches to extract the eigenvalues of a matrix using mainly recurrent neural networks. However, our work focuses on an intrinsically different goal: first, we did not consider the actual interaction matrix of proteins as input data, rather the simpler and less detailed distance matrix; second, our scope is to provide a preliminary example of how to employ DL-based algorithms to extract non-trivial, global structural properties of proteins. Our choice to make use of ENMs relies both on their simplicity and low computational requirements, which allowed us to quickly validate the performance of the CNN.

This validation was carried out through the application of the DAP to our multitask regression problem. In the several fivefold cross-validation processes, the independent folds were built so that a structure and its *reversed* counterpart were included in the same fold. On the other hand, decoys coming from the same protein were allowed to be part of different folds. This results in folds that are not completely independent. In figure 5, we can see an example of the behaviour of training and evaluation losses during the training process on the full training set. The losses have a quite steep decrease during the early stages of the training, where they are almost *coupled*. After a few (approx. 20) epochs the loss on the evaluation set starts oscillating, but it keeps decreasing. These non-negligible oscillations are due partly to the small size of the evaluation set, and partly to the relative lack of robustness of MAPE to small fluctuations.

The result achieved for each eigenvalue in cross-validation and evaluation are reported in figure 6. Results in cross-validation are more accurate than the others: this is reasonable since we decided to include decoys generated from the same protein in different folds. MAPE is a relative performance measure in that it is defined in terms of normalized deviations from the reference value; however, it is not bounded from above; hence, in order to further assess the validity of our predictions, they have to be compared to the ones given by a non-informative model. In figure 7, we can see a comparison between our results on the evaluation set and a non-informative model that always predicts the average value of each eigenvalue in the training set. In our case, we can see that the predictions are considerably more accurate than the ones produced by this non-informative model.

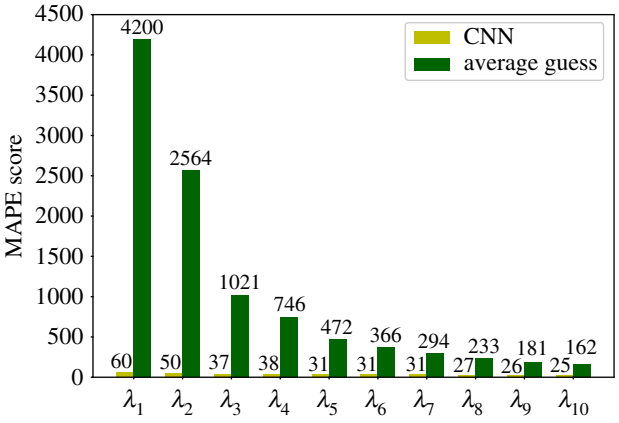

Figure 7. Comparison of the MAPE score as obtained from the CNN trained in the present work (light green bars) and a reference, maximally uninformative random model (dark green bars). The values attained by each model for each eigenvalue are indicated above the bar.

Figure 8 shows the scatter plot of all the eigenvalues in the evaluation set plotted against their predicted values. Since we ran 10 different experiments, we had 20 predictions associated with a single real eigenvalue (sequence-reversed structures share the λ's of the original structures). The almost linear behaviour suggests that the learning model is able to detect with good precision all the eigenvalues even if they range over several orders of magnitude.

The accuracy of the CNN-based approach leaves room for improvement, e.g. through an increased size of the training dataset, a more refined cost function, different parameters and structure of the network etc. However, there is a limitation in the proposed algorithm which is more fundamental than those mentioned, namely the fixed size of the input structures. In fact, the far-reaching objective of employing DL approaches for structural analysis of proteins would be severely limited if only structures with a given number of amino acids could be analysed. A mitigation of this issue comes from the nature of the problem under examination and, as will be illustrated hereafter, opens a novel scenario for the usage of CNNs in the present context.

A few words are in order regarding the computational cost of the CNN in comparison with the exact algorithmic procedure. The time required by the network-based approach to process the entire training set and predict the corresponding eigenvalues is shorter than 5 min on a single-core CPU, while the application of the β-GM to the same dataset requires 25 min on the same platform. It is evident that the amount of information provided by the two approaches, as well as the relative accuracy, are not comparable: while the β-GM produces the full, *exact* spectrum for each molecule (eigenvalues as well as eigenvectors), the CNN can only provide an estimate of the 10 lowest eigenvalues. Nonetheless, even though the computational gain obtained already, in this case, is substantial, one has to bear in mind that the ENM is here taken as reference algorithm precisely because of its velocity and accuracy; on the contrary, we envision applications involving much more time-consuming procedures, e.g. the optimization of complex cost functions [63], for which the speedup can be substantial.

The defining property of low-energy modes of fluctuations is their collective character, which manifests itself in the fact that several residues are displaced in the same direction, with no or very little strain among them. This characteristic lies at the foundation of coarse-graining approaches which aim at identifying large groups of residues behaving as quasi-rigid units [64–72]. It is thus the case that the residues which determine the low energy eigenvalues in ENMs are those few whose distances vary the most, that is, hinge residues. Consequently, it is reasonable to expect that the elimination of a few C_α's from the model would not too drastically impact the value of the computed spectra.

To provide quantitative concreteness to these hypotheses, we fed the CNN, trained to intake 100-residues-long structures, with six proteins of 120 amino acids, 20 of which have been randomly pruned. In figure 9, we show the structure of the selected molecules, which have been chosen from the PDB so as to have some degree of structural variability. These proteins range from very globular (4HNR) to fairly elongated (1BR0) ones, up to a case where a hinge is evident and identifiable already by visual inspection (1E5G). For each of these six molecules, we realized 100 different coarse-grained structures having only 100 amino acids by randomly removing 20 of them. The model set of each protein has been fed to 10 networks, differing only for the initial guess of the hyperparameters. In figure 10, we report the average of the first 10 eigenvalues of each of the six proteins, averaged over the 100 randomly pruned structures and the 10 CNN instances. These eigenvalues are plotted against the value computed by means of the β-GM.

A few observations are in order. In one case, namely 4HNR, there is a perfect overlap between the predictions on the randomly coarse-grained structures and the actual values, with an overall average MAPE equal to 15.8. This molecule is highly globular, which also reflects in the large absolute value of the eigenvalues; hence, it seems that the removal of a relevant fraction of amino acids does not affect the precision of the CNN model. Eigenvalues associated with 2KOK, 2YQD, and 1MEK were predicted with reasonable accuracy, the overall average MAPE being 30.7, 35.5 and 48.3, respectively. These proteins share a medium degree of globularity. The most important deviations from the real eigenvalues appear for 1BR0 and 1E5G, with largely underestimated values; for these molecules, the overall average MAPE equals 73.7 and 58.3, respectively. However, these proteins are at the other extreme of the 'globularity spectrum' with respect to the first, very compact 4HNR. In fact, 1BR0 is a bundle of three quasi-parallel alpha-helices, while 1E5G consists of two identical, independent domains separated by a few interface residues. It is reasonable to expect that in the first case no well-defined hinge region exists, rather each part of the molecule takes part in the low-energy deformation. A random removal of residues thus has an appreciable impact in the calculation of the energetic cost associated with collective motions.

For 1E5G the mechanism is different. This protein possesses a short linker and a relatively small interface connecting two lobes, thus suggesting a rather decoupled dynamics between them. That this is likely the case is made evident by the fact that this molecule features the lowest lowest-energy (sic) eigenvalue among those under examination. Hence, the removal of some residues from those constituting the hinge between the two domains substantially affects the result. In order to verify this hypothesis, we have repeated the CNN-based calculation for 1E5G on a set of 100 pruned structures, which have been obtained by

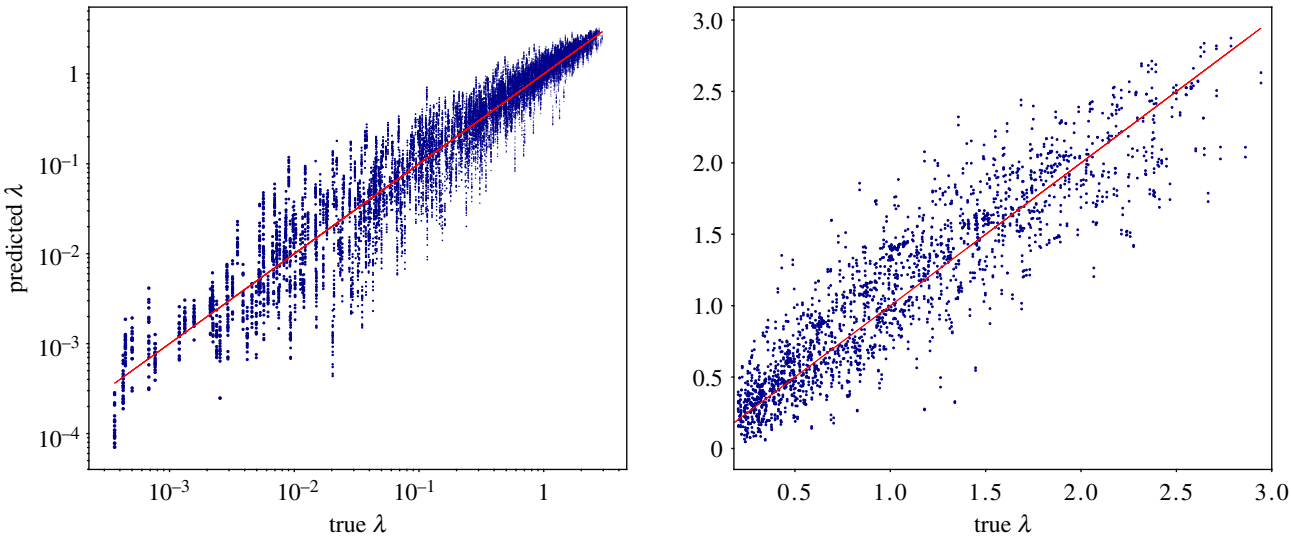

Figure 8. Scatter plot of the eigenvalues predicted by the CNN against the exact ones computed by the β-GM. All lowest 10 non-zero eigenvalues of each protein in the validation set (both 'direct' and 'reverse' orientation of the chain) computed by each of the 10 CNN instances are shown. (*a*) The data are shown in log–log scale. (*b*) Data are reported in linear scale; however, only those eigenvalues having $\lambda > 0.2$ can be seen. (Online version in colour.)

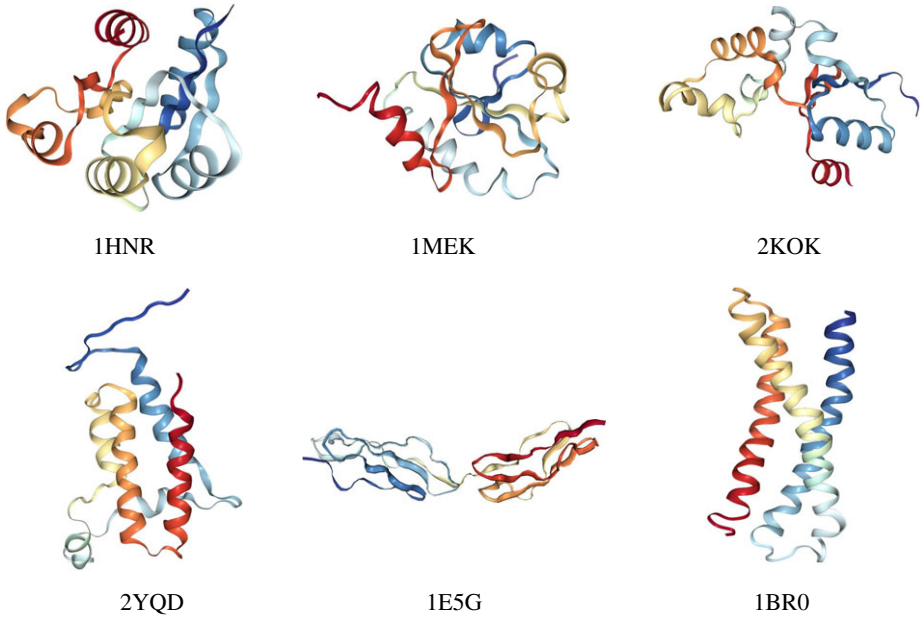

Figure 9. Structures of the six proteins with 120 amino acids employed to test the behaviour of the CNN on molecules larger than those employed for the training. The PDB codes are indicated. The graphical representation has been done using the online tool available on the website of the PDB [73]. (Online version in colour.)

randomly removing 20 amino acids from the crystallographic conformation with the restraint that no residue lying within a 10 Å cut-off from the centre of the linker could be eliminated, as illustrated in figure 11. The effect of this simple criterion can be seen on the data reported in figure 13, which show a small yet appreciable and, most importantly, systematic improvement of the MAPE score when the hinge residues are not selected for removal (*centred sphere CG*) with respect to the completely random selection case (*random CG*).

Finally, to rule out the possibility that this improvement depends on the exclusion of a localized group of CG sites *per se* rather than their particular location, we have performed a further test. Specifically, we have constructed 10 different CG model types in which the 20 exceeding residues have been eliminated outside of a sphere of radius 10 Å whose centre is located on a randomly chosen position of the protein at least 20 Å away from the hinge centre. In plain English, we have performed the same calculation as of the centred sphere CG 10 times, with spheres centred so as to avoid overlap with the one placed in the protein hinge (figure 12). For each model type—i.e. for each location of the exclusion sphere—10 randomized CG models have been produced, and their eigenvalues averaged over specific coarse-graining realization and CNN model.

The result, also visible in figure 13, shows an *increase* of the MAPE score with respect to the random case, that is, the prediction of the CNN worsens with respect to a model where the 20 removed residues have been randomly chosen throughout the structure. This observation consolidates the hypothesis that the network is capable of predicting with sufficient accuracy the low-energy eigenvalues of a protein larger than those it has been trained upon, provided that the exceeding number of sites has been removed; furthermore, and quite intuitively, the prediction improves if the

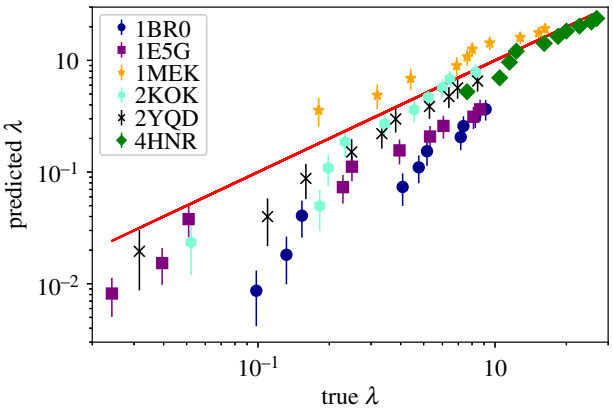

Figure 10. Predicted eigenvalues of the six structures with 120 amino acids against their real value. Each point is the average over 100 random coarse-graining procedures and 10 networks. Error bars indicate the standard deviation. The red line is a guide to the eye. (Online version in colour.)

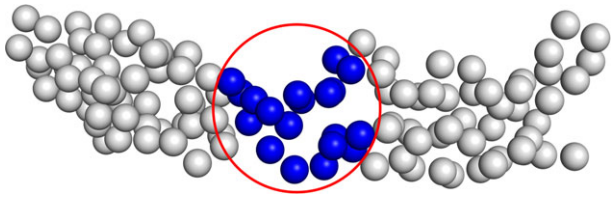

Figure 11. Schematics of the procedure to perform a restrained random removal of the exceeding 20 amino acids from protein 1E5G. Atoms to be eliminated can be selected only outside of the sphere centred on the molecule's hinge and having a 1 nm radius. (Online version in colour.)

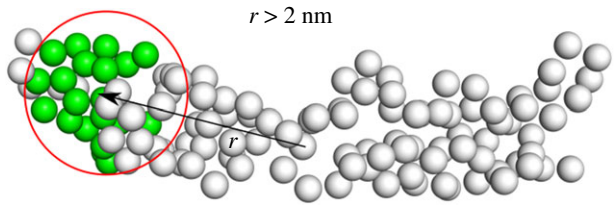

Figure 12. Schematics of the procedure to test whether the improved MAPE values obtained excluding a localized group of residues from removal do not depend on their location. As described in figure 11, the 20 residues to remove can be randomly selected only outside of a sphere of 1 nm radius. The centre of the sphere, however, cannot be localized closer than 2 nm to the point employed for the previous analysis, namely the molecule's sequence centre (i.e. the mechanical hinge). Ten different positions for the sphere are randomly identified, and for each of them, 10 different models where 20 residues have been randomly removed have been constructed. (Online version in colour.)

removed residues do not belong to mechanically relevant parts of the molecule such as motion hinges.

4. Conclusion

The aim of computer-aided modelling of biomolecular systems is to achieve deeper, mechanistic understanding of their function and properties at a level that cannot be accessed by means of experimental or purely theoretical (i.e. mathematical) methods. This approach indeed plays on two sides of a coin: on the one hand, it provides a detailed picture of biological

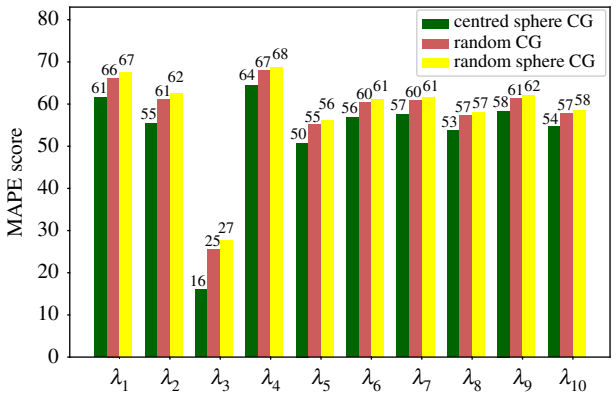

Figure 13. 1E5G: MAPE of predictions on proteins subject to different procedures for the removal of the 20 exceeding residues. *Random CG* stands for randomly coarse-grained structures; *centred sphere CG* refers to the procedure in which central atoms are not removed; and *random sphere CG* describes structures in which the sphere has been placed in regions other than the interface between the domains. (Online version in colour.)

processes at the molecular level, thus enabling the confirmation or falsification of hypotheses and the formulation of theories and models of the finest mechanisms of living matter; on the other hand, it serves as a validation of the currently available representations of the fundamental constituents of cells, ranging from single atoms to entire tissues and organs. Such a workflow is largely *algorithmic and deterministic*, in the sense that it relies on well-defined procedures each step of which is known and understood. An exemplary instrument in this sense is MD.

The alternative strategy, which is gaining further and further attention and interest (as well as success), is machine learning, and DNNs in particular. These computational methods have proven extremely effective in performing those tasks which cannot be easily formulated in a classically *algorithmic* manner, rather they have a fuzzier, more probabilistic character. Nonetheless, a steadily growing level of quantitative accuracy is being reached by DL techniques.

The complementary nature of the two aforementioned approaches is not only extremely appealing but also potentially very powerful, as it is demonstrated for example in the field of bioinformatics, where (big) data processing moves on both tracks simultaneously. In the present work, we have made a first attempt to combine formal, algorithmic models with DL approaches in the context of protein modelling. In particular, it has been our goal to perform, by means of a CNN, the calculation of global properties of protein structures such as the lowest-energy eigenvalues of the most collective modes of fluctuations. The final aim cannot, of course, be that of trivially replacing the simple, extremely effective procedure represented by a matrix inversion by means of a CNN; rather, we explored the possibility of allowing a DL scheme to perform this task with sufficient accuracy as a first, necessary step towards more complex kinds of structural protein analyses. While the calculation of the lowest eigenvalues (as well as the rest of the whole spectrum) of an ENM is immediate and computationally inexpensive in terms of linear algebra, it is not given for granted that a CNN could do it as well. Furthermore, a crucial step in the usage of a CNN (or similar methods) is the pre-processing of the molecular structure in terms of appropriate input variables: the usage of the Coulomb matrix has proven to be a viable choice to this end.

A second, equally relevant outcome of this work is the extension of the network-based eigenvalue prediction network to proteins having a larger number of residues than those employed for the training. The construction of molecular descriptors flexible enough to process proteins of variable length is still an open issue; however, we have shown that the network—trained on 100-residue-long molecules—can provide good estimates of the low-energy eigenvalues of proteins with 120 amino acids provided that the twenty exceeding ones have been neglected. This positive result proves even more pleasant inasmuch as the agreement between predicted and real values varies depending on the specific choice of the removed amino acids, in such a way that mechanically relevant residues emerge as those whose removal determines a worsening of the prediction. The natural consequence of this observation is that, upon appropriate training, DL schemes could be employed in an effective manner not only to compute properties along the lines of reference algorithms but also to extract biologically relevant features of a protein and to provide valuable indication on how to construct simplified, that is, coarse-grained models.

Data accessibility. Raw data are freely and publicly available on the ERC VARIAMOLS project website: http://variamols.physics.unitn.eu.

Competing interests. We declare we have no competing interests.

Funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 758588).

Acknowledgements. We thank Simone Orioli and Flavio Vella for a critical reading of the manuscript and insightful discussion. R.P. gratefully acknowledges the support provided by the Royal Society in the funding and organization of the Theo Murphy international scientific meeting *Multi-resolution simulations of intracellular processes* which took place on 24–25 September 2018 at Chicheley Hall, Buckinghamshire (UK), during which inspiring and fruitful conversations took place.

References

1. Alder BJ, Wainwright TE. 1959 Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459–466. (doi:10.1063/1.1730376)
2. Karplus M. 2002 Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* **35**, 321–323. (doi:10.1021/ar020082r)
3. Takada S. 2012 Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **22**, 130–137. (doi:10.1016/j.sbi.2012.01.010)
4. Noid WG. 2013 *Systematic methods for structurally consistent coarse-grained models*, pp. 487–531. Totowa, NJ: Humana Press.
5. Saunders MG, Voth GA. 2013 Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **42**, 73–93. (doi:10.1146/annurev-biophys-083012-130348)
6. Potestio R, Peter C, Kremer K. 2014 Computer simulations of soft matter: linking the scales. *Entropy* **16**, 4199–4245. (doi:10.3390/e16084199)
7. Pazos F, Chagoyen M. 2016 *Practical protein bioinformatics*, 1st edn. Berlin, Germany: Springer.
8. Wu C, Arighi C, Ross K (eds). 2017 *Protein bioinformatics*, 1st edn. Berlin, Germany: Springer.
9. LeCun Y, Bengio Y, Hinton G. 2015 Deep learning. *Nature* **521**, 436–444. (doi:10.1038/nature14539)
10. Goodfellow I, Bengio Y, Courville A (eds). 2016 *Deep learning*. Cambridge, MA: MIT Press.
11. Marblestone AH, Wayne G, Kording KP. 2016 Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94. (doi:10.3389/fncom.2016.00094)
12. Pak M, Kim S. 2017 A review of deep learning in image recognition, In *2017 4th Int. Conf. on computer applications and information processing technology* (*CAIPT*), pp. 1–3.
13. Krizhevsky A, Sutskever I, Hinton GE. 2012 Imagenet classification with deep convolutional neural networks. In *Proc. 25th Int. Conf. on Neural Information Processing Systems*, vol. 1, pp. 1097–1105. New York, NY: Curran Associates Inc.
14. Dobson CM. 2004 Chemical space and biology. *Nature* **432**, 824–828. (doi:10.1038/nature03192)
15. von Lilienfeld OA. 2012 First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. (http://arxiv.org/abs/1209.5033)
16. Ramakrishnan R, von Lilienfeld OA. 2015 Machine learning, quantum mechanics, and chemical compound space. (http://arxiv.org/abs/1510.07512)
17. Todeschini R, Consonni V. 2000 *Handbook of molecular descriptors*. Weinheim, Germany: WileyVCH.
18. Rogers D, Hahn M. 2010 Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754. (doi:10.1021/ci100050t)
19. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. 2016 Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608. (doi:10.1007/s10822-016-9938-8)
20. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W. 2008 Mold2, molecular descriptors from 2d structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **48**, 1337–1344. (doi:10.1021/ci800038f)
21. Vračko M. 2015 Mathematical (structural) descriptors in QSAR: applications in drug design and environmental toxicology. In *Advances in mathematical chemistry and applications* (eds SC Basak, G Restrepo, JL Villaveces), pp. 222–250. Oxford, UK: Bentham Science Publishers.
22. Imbalzano G, Anelli A, Giofré D, Klees S, Behler J, Ceriotti M. 2018 Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730. (doi:10.1063/1.5024611)
23. Grisafi A, Fabrizio A, Meyer B, Wilkins DM, Corminboeuf C, Ceriotti M. 2019 Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64. (doi:10.1021/acscentsci.8b00551)
24. Bohacek RS, McMartin C, Guida WC. 1996 The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50. (doi:10.1002/(ISSN)1098-1128)
25. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. 2001 Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26. (doi:10.1016/S0169-409X(00)00129-0)
26. Geiger P, Dellago C. 2013 Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.* **139**, 164105. (doi:10.1063/1.4825111)
27. Feinberg EN *et al.* 2018 Potentialnet for molecular property prediction. *ACS Cent. Sci.* **4**, 1520–1530. (doi:10.1021/acscentsci.8b00507)
28. Wehmeyer C, Noé F. 2018 Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **148**, 241703. (doi:10.1063/1.5011399)
29. Wang S, Peng J, Ma J, Xu J. 2016 Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* **6**, 18962. (doi:10.1038/srep18962)
30. Wang S, Sun S, Li Z, Zhang R, Xu J. 2017 Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324. (doi:10.1371/journal.pcbi.1005324)
31. Lemke T, Peter C. 2017 Neural network based prediction of conformational free energies—a new route toward coarse-grained simulation models. *J. Chem. Theory Comput.* **13**, 6213–6221. (doi:10.1021/acs.jctc.7b00864)
32. Zhang L, Han J, Wang H, Car R, Weinan E. 2018 DeePCG: constructing coarse-grained models via deep neural networks. *J. Chem. Phys.* **149**, 034101. (doi:10.1063/1.5027645)
33. Tirion MM. 1996 Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908. (doi:10.1103/PhysRevLett.77.1905)

34. Micheletti C, Carloni P, Maritan A. 2004 Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins: Struct. Funct. Bioinf.* **55**, 635–645. (doi:10.1002/prot.20049)
35. Cireşan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J. 2011 Flexible, high performance convolutional neural networks for image classification. In *Proc. 22nd Int. Joint Conf. on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011,* pp. 1237–1242. Palo Alto, CA: AAAI Press.
36. Chellappa Z. 1988 Computation of optical flow using a neural network. In *IEEE 1988 Int. Conf. on Neural Networks, San Diego, CA, USA, 24–27 July 1988,* vol. 2, pp. 71–78. (doi:10.1109/ICNN.1988.23914)
37. Anderson JA, Lorenz CD, Travesset A. 2008 General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **227**, 5342–5359. (doi:10.1016/j.jcp.2008.01.047)
38. Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. 2007 Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* **28**, 2618–2640. (doi:10.1002/(ISSN)1096-987X)
39. Fan Z, Chen W, Vierimaa V, Harju A. 2017 Efficient molecular dynamics simulations with many-body potentials on graphics processing units. *Comput. Phys. Commun.* **218**, 10–16. (doi:10.1016/j.cpc.2017.05.003)
40. Shaw DE *et al.* 2009 Millisecond-scale molecular dynamics simulations on Anton. In *Proc. Conf. on High Performance Computing Networking, Storage and Analysis, Portland, OR, USA, 14–20 November 2009*, article 39. New York, NY: ACM. (doi:10.1145/1654059.1654099)
41. Delarue M, Sanejouand Y-H. 2002 Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.* **320**, 1011–1024. (doi:10.1016/S0022-2836(02)00562-4)
42. Tama F, Brooks CL. 2002 The mechanism and pathway of ph induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* **318**, 733–747. (doi:10.1016/S0022-2836(02)00135-3)
43. Valadié H, Lacapčre J, Sanejouand Y-H, Etchebest C. 2003 Dynamical properties of the mscl of *Escherichia coli*: a normal mode analysis. *J. Mol. Biol.* **332**, 657–674. (doi:10.1016/S0022-2836(03)00851-9)
44. Tama F, Valle M, Frank J, Brooks CL. 2003 Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl Acad. Sci. USA* **100**, 9319–9323. (doi:10.1073/pnas.1632476100)
45. Park BH, Levitt M. 1996 Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392. (doi:10.1006/jmbi.1996.0256)
46. Crecca CR, Roitberg AE. 2008 Using distances between α-carbons to predict protein structure. *Int. J. Quantum Chem.* **108**, 2782–2792. (doi:10.1002/qua.v108:15)
47. Amadei A, Linssen ABM, Berendsen HJC. 1993 Essential dynamics of proteins. *Proteins: Struct. Funct. Bioinf.* **17**, 412–425. (doi:10.1002/(ISSN)1097-0134)
48. Amadei A, Ceruso MA, Di Nola A. 1999 On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct. Funct. Bioinf.* **36**, 419–424. (doi:10.1002/(ISSN)1097-0134)
49. Daidone I, Amadei A. 2012 Essential dynamics: foundation and applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 762–770. (doi:10.1002/wcms.1099)
50. Diggins P, Liu C, Deserno M, Potestio R. 2019 Optimal coarse-grained site selection in elastic network models of biomolecules. *J. Chem. Theory Comput.* **15**, 648–664. (doi:10.1021/acs.jctc.8b00654)
51. Behler J, Parrinello M. 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401. (doi:10.1103/PhysRevLett.98.146401)
52. Behler J. 2014 Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.: Condens. Matter* **26**, 183001. (doi:10.1088/0953-8984/26/18/183001)
53. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014 Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958.
54. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R. 2012 Improving neural networks by preventing co-adaptation of feature detectors. (http://arxiv.org/abs/1207.0580)
55. Chollet F *et al.* 2015 Keras. https://keras.io.
56. Abadi M *et al.* 2015 TensorFlow: large-scale machine learning on heterogeneous systems, Software available from tensorflow.org.
57. Duchi J, Hazan E, Singer Y. 2011 Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159.
58. de Myttenaere A, Golden B, Grand BL, Rossi F. 2016 Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48. (doi:10.1016/j.neucom.2015.12.114)
59. MAQC. Consortium. 2010 The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838. (doi:10.1038/nbt.1665)
60. SEQC/MAQC-III. Consortium. 2014 A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.* **32**, 903–914. (doi:10.1038/nbt.2957)
61. Yi Z, Fu Y, Tang HJ. 2004 Neural networks based approach for computing eigenvectors and eigenvalues of symmetric matrix. *Comput. Math. Appl.* **47**, 1155–1164. (doi:10.1016/S0898-1221(04)90110-1)
62. Finol D, Lu Y, Mahadevan V, Srivastava A. 2018 Deep convolutional neural networks for eigenvalue problems in mechanics. (http://arxiv.org/abs/1801.05733)
63. Diggins P, Liu C, Deserno M, Potestio R. 2019 Optimal coarse-grained site selection in elastic network models of biomolecules. *J. Chem. Theory Comput.* **15**, 648–664. (doi:10.1021/acs.jctc.8b00654)
64. Hinsen K. 1998 Analysis of domain motions by approximate normal mode calculations. *Proteins* **33**, 417–429. (doi:10.1002/(ISSN)1097-0134)
65. Golhlke H, Thorpe MF. 2006 A natural coarse graining for simulating large biomolecular motion. *Biophys. J.* **91**, 2115–2120. (doi:10.1529/biophysj.106.083568)
66. Zhang Z, Lu L, Noid WG, Krishna V, Pfaendtner J, Voth GA. 2008 A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.* **95**, 5073–5083. (doi:10.1529/biophysj.108.139626)
67. Zhang Z, Pfaendtner J, Grafmüller A, Voth GA. 2009 Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys. J.* **97**, 2327–2337. (doi:10.1016/j.bpj.2009.08.007)
68. Potestio R, Pontiggia F, Micheletti C. 2009 Coarse-grained description of proteins' internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.* **96**, 4993–5002. (doi:10.1016/j.bpj.2009.03.051)
69. Aleksiev T, Potestio R, Pontiggia F, Cozzini S, Micheletti C. 2009 Pisqrd: a web server for decomposing proteins into quasi-rigid dynamical domains. *Bioinformatics* **25**, 2743–2744. (doi:10.1093/bioinformatics/btp512)
70. Zhang Z, Voth GA. 2010 Coarse-grained representations of large biomolecular complexes from low-resolution structural data. *J. Chem. Theory Comput.* **6**, 2990–3002. (doi:10.1021/ct100374a)
71. Sinitskiy AV, Saunders MG, Voth GA. 2012 Optimal number of coarse-grained sites in different components of large biomolecular complexes. *J. Phys. Chem. B* **116**, 8363–8374. (doi:10.1021/jp2108895)
72. Polles G, Indelicato G, Potestio R, Cermelli P, Twarock R, Micheletti C. 2013 Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition. *PLoS Comput. Biol.* **9**, e1003331. (doi:10.1371/journal.pcbi.1003331)
73. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000 The protein data bank. *Nucleic Acids Res.* **28**, 235–242. (doi:10.1093/nar/28.1.235)