The AAPS Journal (2026) 28:87
https://doi.org/10.1208/s12248-026-01232-z

RESEARCH ARTICLE

Predicting Aberrant Fc-fusion Protein Pharmacokinetics from *In Silico* Structural Properties and Physiologically Based Pharmacokinetic (PBPK) Modeling

Danilo Tomasoni[1] · Alessio Paris[1] · Roberto Visintainer[1] · Kevin D. Cook[2] · Aochiu Chen[3] · Isabel Figueroa[2] · Veena A. Thomas[2] · Luca Marchetti[1,4]

Received: 28 October 2025 / Accepted: 4 March 2026

Abstract

The fusion of therapeutic proteins to the Fc domain of monoclonal antibodies (mAbs) generally improves the proteins' pharmacokinetic (PK) characteristics, extending *in vivo* half-lives due to the binding of the Fc domain to the FcRn receptor. Yet, several of these Fc-fusion biologics have been observed to have unexpected rapid clearance associated with non-specific off-target binding. Variability in non-specific clearance is often challenging to predict, not well understood, and ultimately can delay the drug development process. In this investigation, we present a computational approach leveraging *in silico* protein structural properties to extend a physiologically based pharmacokinetic (PBPK) model of mAbs validated on *in vivo* plasma PK profiles in mice. Selected model parameters affecting protein half-life have been scaled by analytical functions of a panel of calculated *in silico* protein properties identified by a novel and ad hoc symbolic regression procedure. The resulting extended model has been successfully validated against an independent set of protein plasma PKs, indicating that it can generalize to novel biologics of the same class. Moreover, the extended PBPK model has a median absolute average fold error (AAFE) of 1.18 (min = 1.09; max = 1.51), where values less than 2 typically indicate a good fit. The results enable the de-risking of aberrant PK behaviors, ultimately leading to the selection of Fc-fusion proteins with increased therapeutic value for patients.

Keywords Fc-fusion proteins · monoclonal antibodies · ordinary differential equations · physiologically based pharmacokinetic modeling · symbolic regression

Introduction

With around one hundred approved products, monoclonal antibody (mAb) derived drugs represent a relevant share of the current pharmaceutical research and development (1). The success of immunoglobulin (IgG) derived biotherapeutics resides in different factors: high target specificity, Fc-modulated functionality, and favorable pharmacokinetic (PK), as reflected in its slow clearance and long half-life (2).

Among these products, a relevant group is represented by Fc-fusion proteins synthesized by fusing the fragment crystallizable (Fc) domain of IgG to a biologically active protein or peptide (3, 4). Fc-fusion engineering strategy mitigates the inherently rapid systemic clearance of these biologically active proteins by conferring resistance to proteolytic catabolism. The engineered Fc-domain extends the circulation time of the therapeutics in the bloodstream by binding to the neonatal Fc receptor (FcRn) in the endosomes

✉ Veena A. Thomas
vthoma03@amgen.com

✉ Luca Marchetti
marchetti@cosbi.eu; luca.marchetti@unitn.it

1 Fondazione The Microsoft Research – University of Trento Centre for Computational and Systems Biology (COSBI), Rovereto, Italy

2 Pharmacokinetics and Drug Metabolism-Bioanalytical Sciences, Amgen, South San Francisco, California, USA

3 Biologics, Amgen, South San Francisco, California, USA

4 Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy

Published online: 07 April 2026

of endothelial cells, thus protecting the Fc-fusion protein from lysosomal degradation.

However, significant inter-biologic variability in the plasma clearance among Fc-fusion proteins is often observed (5). The primary elimination mechanisms of Fc-based biologics are antigen-mediated endocytosis, intracellular catabolism via non-specific fluid-phase endocytosis, and in rare cases anti-drug immunogenic response. Antigen-mediated clearance is often dose-dependent and can be mitigated by increasing the dose. Similarly, several assays are in place to mitigate risks associated with anti-drug antibodies. Utilization of immunocompromised mice or truncation of the concentration–time profile can be leveraged to address the effect of anti-drug antibodies. In contrast, unexpected fast clearance associated with the non-specific off-target binding is difficult to predict. The source of non-specific clearance is often unknown and non-saturable. Unexpected fast clearance of Fc-based biologics is identified in rodents or non-human primates' PK studies, often in the late optimization stage of drug development. This may result in significant attrition and critical delays in drug development. This investigation is a comprehensive evaluation of the predictive utility of *in silico* molecular metrics to *a priori* predict the developability of a panel of Fc-fusion proteins. The results presented in this study could accelerate the timeline towards the first-in-human clinical trials while complying with the principles of replacement, reduction, and refinement of animals used in research of the "3Rs alternatives" and the Food and Drug Administration Modernization Act 2.0. (6).

Previous works have identified potential factors affecting the non-specific binding of mAbs or mAb-derived therapeutics, such as isoelectric point (7), charged patches (8, 9), and hydrophobicity (10). Concurrently, several methodological frameworks have explored quantitative prediction of preclinical or clinical clearance using *in vitro* or *in silico* metrics (11, 12). Among the available approaches, we employed a physiologically based pharmacokinetic (PBPK) platform. Recognized as a 'gold standard', PBPK models provide a comprehensive mathematical representation of critical physiological processes governing antibody uptake, distribution, and elimination (13).

We hypothesize that the PBPK model built on *in silico* molecular metrics of Fc-fusion proteins will de-risk aberrant PK behaviors stemming from non-specific binding. In this work, we modified an existing PBPK framework by Shah and Betts (14), and incorporated *in silico* physicochemical properties of a panel of Fc-fusion proteins as covariates of physiological processes including the non-specific endosomal uptake of the individual proteins, and the recycling from the lymphatic flow to the bloodstream. We utilized a dataset from a panel of Fc-fusion proteins in which a series of five *de novo* generated scaffolds was each modified to generate variant series that differed in size, shape, and physicochemical properties (15). The plasma concentration–time profiles of this panel of Fc-fusion proteins were assessed in mouse and importantly these proteins were murine non-cross reactive, minimizing the likelihood of target-mediated clearance due to the absence of target engagement in host species. For each Fc-fusion protein, relevant *in silico* characteristics were computed from the simulated folded structure of the protein and were included in the PBPK model as covariates of two factors, labeled F1 and F2, scaling model parameters associated with the aforementioned physiological processes. We devised an ad hoc symbolic regression procedure with cross-validation that allowed us to determine analytical formulas for F1 and F2 in terms of relevant *in silico* properties of proteins from the training set ($n = 34$). Subsequently, the extended model was evaluated according to the accuracy in reproducing or predicting the PK profiles of the Fc-fusion proteins in the validation set ($n = 7$).

Materials and Methods

Calculation of *In Silico* Properties

A FASTA file of designed sequences for structure prediction was generated for each protein. Only the fused protein, not the Fc, was considered to reduce the computational power required. The folded structure of each protein was designed using AlphaFold (16) without multiple sequence alignment or template information to avoid biases. The best-folded structure was selected considering different factors (predicted local distance difference test (pLDDT) score, root-mean-square deviation to the template molecule, and side chain clashes) and prepared for simulation in MOE (Molecular Operating Environment) (17), a software package developed by the Chemical Computing Group. Simulations were run using the force field Amber10:EHT. For each protein model, 100 ensembles were generated using LowModeMD (18) within a pH range of 6.4 to 8.4. The final output is the average of the properties for each protein across all 100 ensembles.

In Silico Properties Selection

Each Fc fusion protein has a list of 55 physicochemical properties. Unfortunately, the potential algebraic combinations of the properties increase exponentially with the number of properties considered, and property collinearities are known to harm the stability of linear models (19).

To mitigate this risk, we employed a three-step approach: firstly, an automated procedure (19) based on Ward's minimum variance criterion (20) was used to compute the best number of clusters in which the properties can be classified. The best number of clusters resulted to be 15 (out of

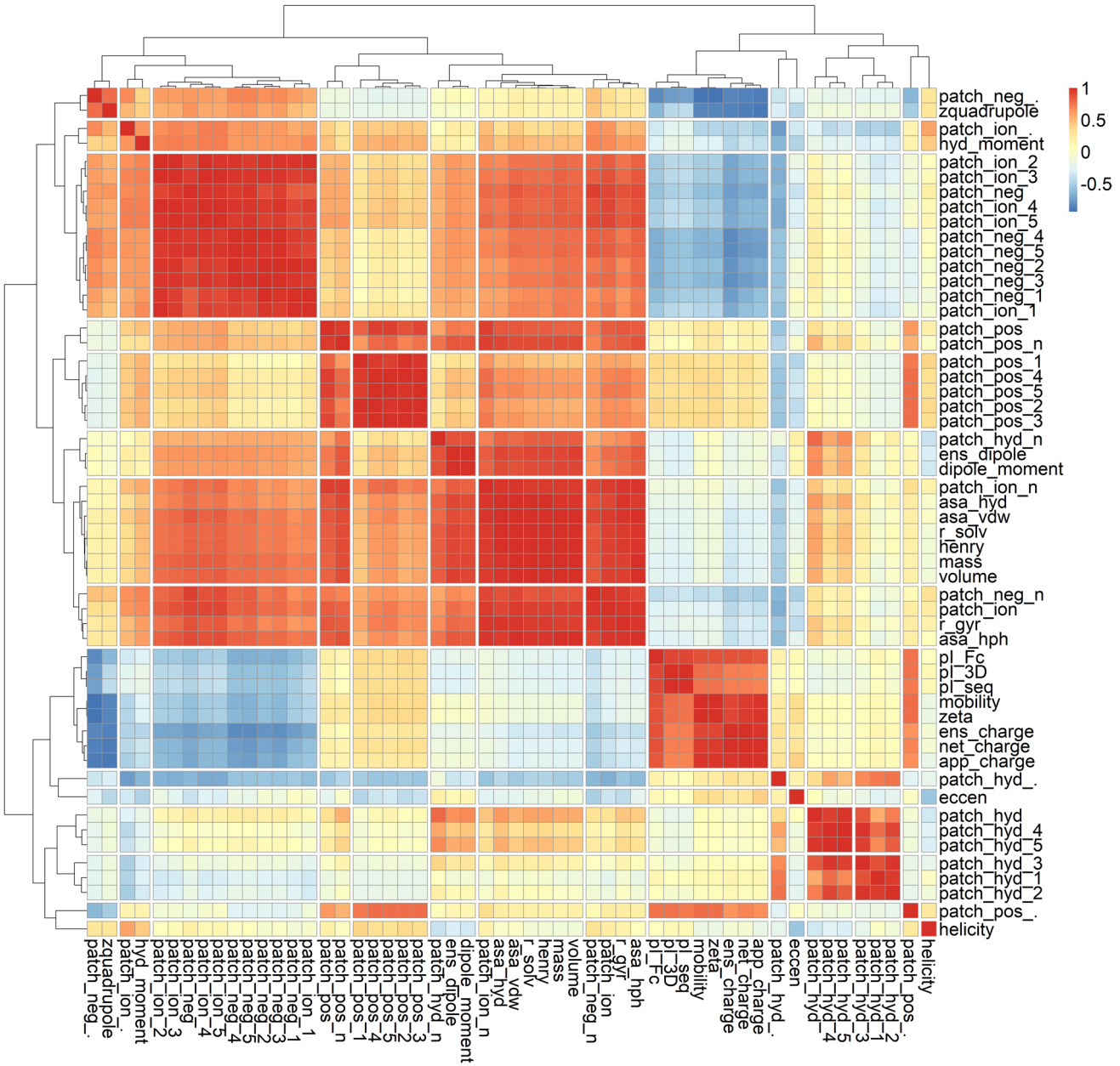

Fig. 1 Hierarchical clustering based on correlation between physicochemical properties of Fc-fusion proteins

55 covariates). Then, we computed a hierarchical clustering based on the correlation matrix between the properties (see Fig. 1), setting the threshold of the clustering in such a way that the final number of clusters is exactly 15, as calculated by the previous analysis. Finally, for each cluster, we selected the medoid, that is, the point with minimum distance to all other points in the cluster, using the R package monoClust. The fifteen medoids were used for the subsequent phase of identifying the formulas.

Identification of Formulas for F1 and F2

The problem of finding a mathematical relation that links input and output data is well-known and studied. Many solutions have been proposed in the literature, with the most popular being based on genetic algorithms (22). Recently, the rise of deep learning (DL) has inspired new DL-based approaches as well (23, 24).

The difficulties primarily lie in the dimension of the search space, which, in the most general case, is infinite, and in the size of the available dataset, which is often very limited.

When the dataset is big enough, DL-based approaches tend to provide better results, but in our scenario and in many other scientific settings, the number of experiments is limited due to ethical and economic reasons.

The current work addresses these issues by introducing the concept of bounded search, which draws inspiration from the Occam's razor principle, which states that simpler models should be preferred over more complex ones. This means that we put some realistic bounds on the formula structure and length and start our exploration from simpler formulas, with one or zero interactions among input features (a, a*b…), a few types of non-linear transformations (1/a, log10(a)…), and a small number of terms (a + b, a + b + c…). Given these bounds, the search space is limited and relatively small to explore. Only if we cannot find a satisfactory formula, we relax the bounds to allow for more interactions, more non-linear transformations, or more terms.

Formula Evaluation

To choose the best formula available, we compute a performance score (S) as follows. Our formula evaluations are guided by N repetitions, with order randomization, of the K-fold cross-validation (25). We consider the results in terms of percent relative error PE = 100*|obs-pred|/obs (median, max and inter-quartile-range), correlation, R^2, and Cooks distance (max).

The final performance score assigned to the formulas is given by Eq. (1):

$$S = \frac{\left(\frac{R^2}{0.4}\right)^2 * sign\left(R^2\right)}{e^{10*PE*\frac{len}{\max len}}}, \quad (1)$$

where R^2 is the coefficient of determination, 0.4 is an empirical threshold, sign() returns 1 if its argument is positive, −1 otherwise, PE is the percentage relative error, len is the number of terms in the formula and maxlen is the current bound on the number of terms the formula is allowed to have.

Formula Generation

To explore the bounded search space, we employed two complementary strategies, random sampling and genetic optimization, briefly described below:

- **Random sampling** efficiently generates random samples from the search space by exploiting the concept of combination rank (26): a unique ID associated with each possible combination of formula terms of a fixed length. By randomly sampling an ID and inverting the rank, we can efficiently sample formulas without generating the whole set of combinations in advance.
- **Genetic optimization** (27) first encodes the bounded set of possible terms, determined by the maximum number of interactions and non-linear transformations, in a binary array that can be fed to a genetic algorithm. The core of this method is a custom fitness function that equals Eq. (1) if the generated formula is within actual bounds, -infinite otherwise.

Both methods memoize (28) results between subsequent runs to avoid evaluating the same formula twice and produce a table of results with the same column structure. In this way results from both methods can be analyzed and joined together.

Further, due to the independent nature of each formula, multiple searches with different starting points (random seeds) and algorithms can be run in parallel to exploit modern CPU parallelism and reduce exploration time.

We provide an implementation of the ensemble method described in the R package "SymbolicR" that we release through the GitHub on-line platform (29).

Results

To link the physicochemical properties of Fc-fusion proteins to their pharmacokinetic (PK), we considered a set of 41 d*e novo* designed proteins (15), identified by an identification number, with different scaffolds, sizes (from 6 to 20 kDa), charges, and hydrophobicity. PK studies in mice (15) measured the plasma concentration profiles up to 168 h after intravenously administering 2 mg/kg of proteins. Of the 41 Fc-fusion proteins considered in this work, 35 proteins were generated in a bivalently formatted construct, with 2 copies of the fusion protein. The other 6 Fc-fusions were produced in a monovalent format, consisting of a single fusion protein. A visual inspection of the profiles indicates that all products deviate to different extents from the typical mAbs clearance, with remarkable differences among proteins and some atypical drops or shapes. These differences are reflected in the wide range of measured $AUC_{(0\text{-last})}$ (min:$1.58*10^3$ ng*h/ml, max: $1.31*10^6$ ng*h/ml), which corresponds to a high variability of the plasma clearance of the Fc-fusion proteins.

To address the challenge of predicting the protein PK, we introduced structural modifications to the foundational PBPK model structure proposed by Shah and Betts (14) using factors dependent on the physicochemical properties of the proteins.

To this goal, a comprehensive dataset of 55 properties was computed *in silico* as the average of an ensemble of folded protein structures at different pH values. The set encompasses structural properties (mass, volume, radius,

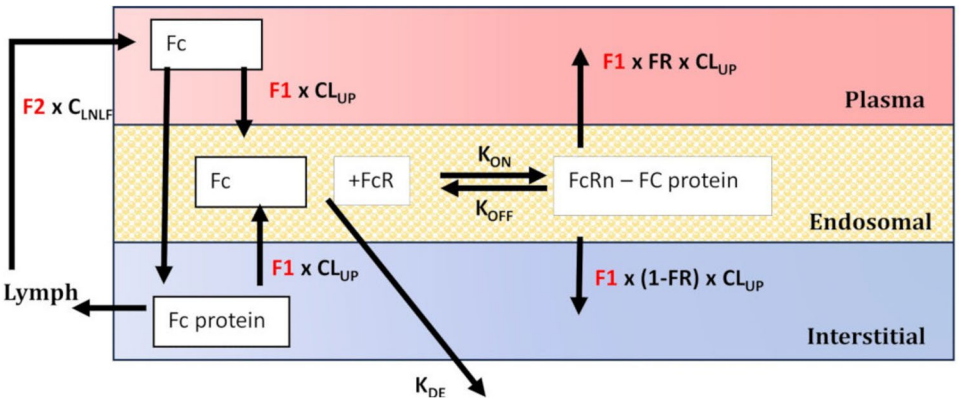

Fig. 2 General schematic of PBPK tissue sub-compartments for Fc protein. The diagram is a modified version of the organ sub-compartment of the PBPK model by Shah and Betts (14). Highlighted in red are the scaling factors F1 and F2 considered in this work, scaling the non-specific uptake CL_{UP} and the lymph-to-blood rate C_{LNLF}, respectively. The distribution of Fc proteins across organ sites remains the same as in the original model (see Fig. 1 in (14))

etc.) along with charge properties (net charge, positive or negative patches, etc.) and hydrophobicity (hydrophobic patches, etc.). A concise description of each *in silico* property is reported in Supplementary File S1, Table S1. A hierarchical clustering algorithm (30) allowed us to identify clusters of highly correlated properties, as reported in

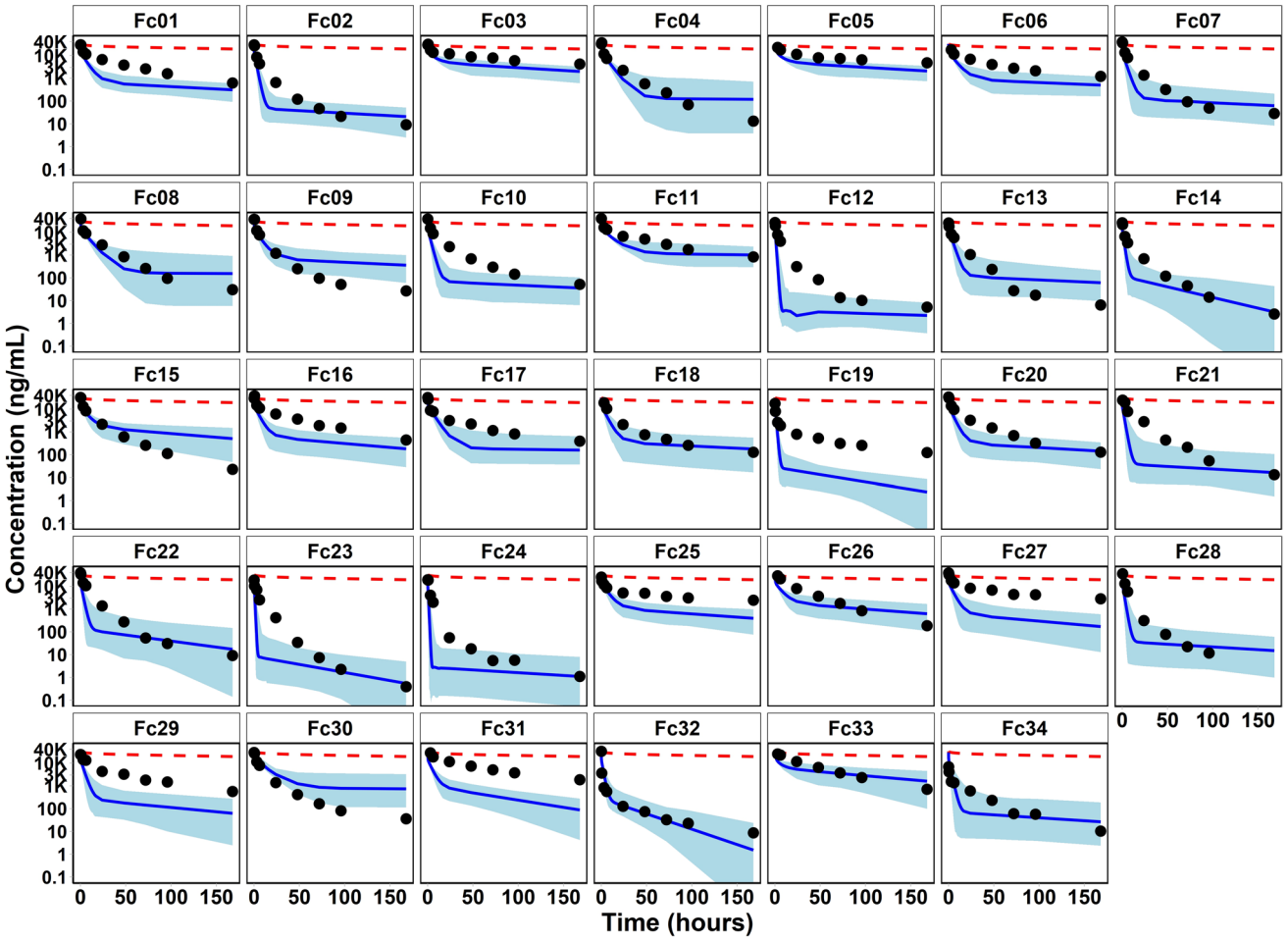

Fig. 3 Measured (black dots) and simulated (blue lines and bands) pharmacokinetic time series in plasma for the proteins of the training set. Each protein is identified by the ID number in the top part of each panel. Simulations were run using the modified PBPK model by Shah and Betts (14) with the scaling factors F1 and F2 computed for each protein as in Eqs. (2, 3, 4, 5 and 6). The bands were simulated with a 20% variation of the coefficients of the formulas for F1 and F2. The simulation of a non-specific mAb product using the original model is also reported (red dashed lines)

a previous subsection "Identification of formulas for F1 and F2". For each cluster, only one representative property, selected with the medoid method, was kept for the following modeling steps, leaving us with a reduced set of 15 properties (app_charge, asa_vdw, eccen, ens_dipole, helicity, patch_hyd_%, patch_hyd_2, patch_hyd_5, patch_ion, patch_ion_%, patch_ion_3, patch_pos, patch_pos_%, patch_pos_3, pI_Fc, we refer to Supplementary Table S1 in Supplementary File S1 for their description). This preliminary identification also reduced the number of property combinations to be explored in the following regression procedure. After selecting the *in silico* properties, the proteins were divided into two sets of 34 and 7 proteins, respectively, to train and validate the modified model. The final step was to normalize and standardize the *in silico* property values of both groups, using the means and standard deviations computed for the training set. Since training and validation data arrived in different batches at different times, it was not possible to assign the proteins randomizing with respect to the *in silico* property distributions. Yet, we checked at a later time that the two sets have overlapping distributions for all 15 properties, despite the small size of the validation set.

After this preliminary identification of potentially relevant properties, we modified the PBPK model of Shah and Betts, introducing two additional factors, F1 and F2, scaling pinocytosis rates CL_{UP} and lymph-to-blood return C_{LNLF} respectively, as represented in the diagram of Fig. 2. This pair of modified parameters was the best combination in terms of fit accuracy among the numerous different pairs of parameters that were preliminarily tested to fit the data. In choosing which parameters to scale among different pairs, CL_{UP} was considered in all cases, while C_{LNLF} was selected among the model parameters showing a sensitivity on the PK plasma protein profile that could plausibly be affected by a protein-dependent variability (C_{LNLF}, FR, $sigma_V$, L k_{deg}, k_{on}, k_{off}) (14).

The effect of scaling CL_{UP} primarily manifests as a change in the slope of the PK time profile at later time-points. On the other hand, scaling C_{LNLF} predominantly influences the initial distribution phase, attenuating the magnitude of the initial concentration drop, and producing a more gradual transition between distribution and elimination phase, particularly evident in log-scale plots (Fig. 3 and Fig. 4).

F1 and F2 were estimated using the weighted least squares cost function (weights proportional to the observations squared) for the multi-start runs of the Nelder-Mead (NM) optimization algorithm (31). The computed values of F1 range from 26.64 to 41462.53, while F2 goes from $2.01*10^{-9}$ to 64.5. The relative least square error of the fits has median 0.75 (min = 0.05, max = 6.96), and the median percentage error has median value 23% (min = 4.5% (id Fc03), max = 88.4% (id Fc34)).

We used these estimates to determine formulas for F1 and F2 following the symbolic regression procedure explained in the Materials and Methods section. To be more precise, given that the values of the two parameters span different orders of magnitude, we identified formulas for the logarithms of F1 and F2. The procedure followed a two-step sequential refinement strategy: first, we determined a regression formula for F1. Subsequently, we incorporated the formula into the PBPK model and performed another round of fits, further refining the coefficients of the regression formula for F1 (labeled β_1 to β_4) and the individual values of F2. Then, we identified a regression formula for the updated F2 values using the symbolic regression method. Finally, we included the formula for F2 in the model and fitted the data again, refining all the formula coefficients (from $\beta_1$to β_7). From a computational standpoint, this iterative refinement strategy offers greater efficiency compared to the simultaneous ab initio estimation of F1 and F2, as it reduces the dimensionality of the optimization space and facilitates more stable convergence. It also guaranteed better average predictive performance among all Fc-fusion proteins considered.

The explicit expressions for the logarithm (base 10) of F1 and F2 are reported in Eqs. (2 and 3). The constants A, B, and C, whose expressions are reported in Eqs. (4, 5, and 6), are introduced to avoid negative arguments in the logarithms, adding a fixed positive shift.

Table I Mean and Standard Deviation (SD) of the *in silico* properties included in Eqs. (2, 3, 4, 5 and 6) for F1 and F2, computed for the proteins in the training set. These values are used to standardize the properties of both the training and validation sets

	app_charge	patch_hyd_2	patch_hyd_5	patch_pos	patch_pos_%	patch_pos_3	pI_Fc
Mean (SD)	−1.99(3.36)	256.87(93.38)	380.36(152.68)	516.63(353.08)	10.88(4.26)	268.52(112.91)	6.73(0.98)

Table II Values of the coefficients of the predictive Eqs. (2, 3) for the scaling factors F1 and F2

Parameter	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Value	3.53	0.85	2.40	0.40	−5.97	−0.36	0.34

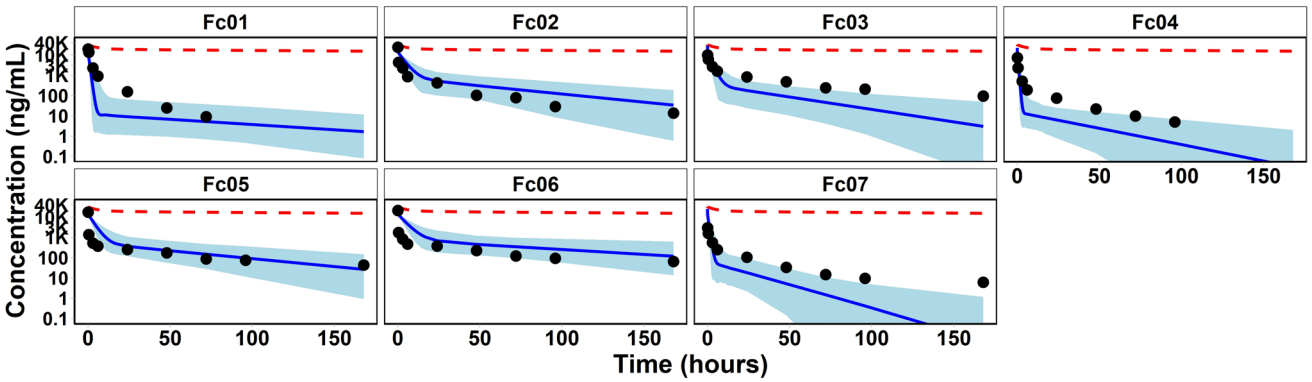

Fig. 4 Measured (black dots) and simulated (blue lines and bands) pharmacokinetic time series in plasma for the proteins of the validation set. Each protein is identified by the ID number in the top part of each panel. Simulations were run using the modified PBPK model by Shah and Betts (14) with the scaling factors F1 and F2 computed for each protein as in Eqs. (2, 3, 4, 5 and 6). The bands were simulated with a 20% variation of the coefficients of the formulas for F1 and F2. The simulation of a non-specific mAb product using the original model is also reported (red dashed lines)

$$\log_{10}(F1) = \beta_1 + \beta_2 \log_{10}\left(A + app_{charge} * pI_{FC}\right) + \beta_3 \log_{10}\left(B + patch_{pos_3}\right) + \beta_4 patch_{pos} * pI_{Fc} \quad (2)$$

$$\log_{10}(F2) = \beta_5 + \beta_6 / \left(C + patch_{hyd_2}\right) + \beta_7 patch_{hyd_5} * patch_{pos_{\%}} \quad (3)$$

$$A = 0.1 + |\min\left(app_{charge} * pI_{FC}\right)| \quad (4)$$

$$B = 0.1 + |\min\left(patch_{pos_3}\right)| \quad (5)$$

$$C = 0.1 + |\min\left(patch_{Hyd_2^2}\right)| \quad (6)$$

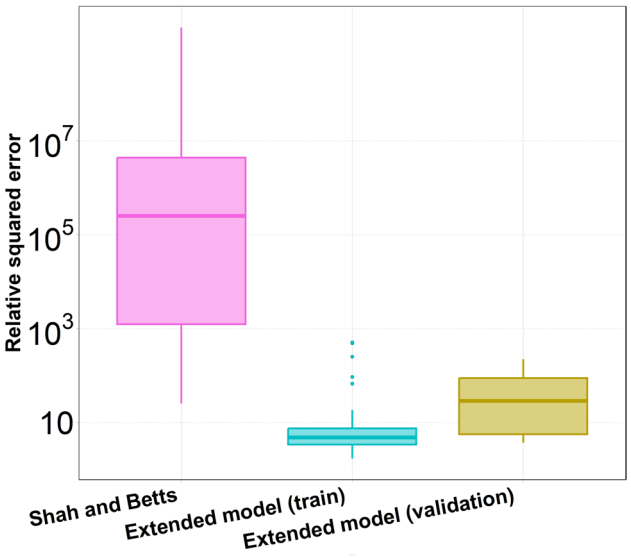

Fig. 5 Relative squared error of simulated time series with respect to measured PK in plasma, as computed for all proteins using the original Shah and Betts (14) PBPK model (left box), and using the modified model with F1 and F2 as computed from Eqs. (2, 3, 4, 5 and 6) for the proteins of the training (central box) and validation set (right box)

The final formulas have a contribution from a small number of covariates: app_charge, patch_hyd_2, patch_hyd_5, patch_pos, patch_pos_%, patch_pos_3, pI_Fc. Their mean and standard deviation, used to standardize the protein values, are reported in Table I, while the values of the formula coefficients are reported in Table II.

The results of the final fits for the training set are reported in Fig. 3, superimposed on the raw data and the original Shah model predictions. We also simulated the PK for each protein with a standard 20% perturbation of the coefficients of the two formulas (light blue bands in Fig. 3). This verification is essential to assess the stability of our extended model because the coefficients contribute to F1 and F2 exponentially, and this could lead to uncontrolled changes in the predicted PK profiles even with slight variations in the coefficients.

To validate the model, we considered 7 proteins excluded from the model training and kept for this purpose. In this case, we normalized and standardized the values of the protein properties according to the mean and standard deviation computed on the training proteins only. We simulated the PK profiles according to the modified PBPK model. The results are reported in Fig. 4, with the 20% perturbation bands of the coefficients.

To assess quantitatively the enhanced predictive performance of the extended model, we compared the relative square errors of simulations run using the original

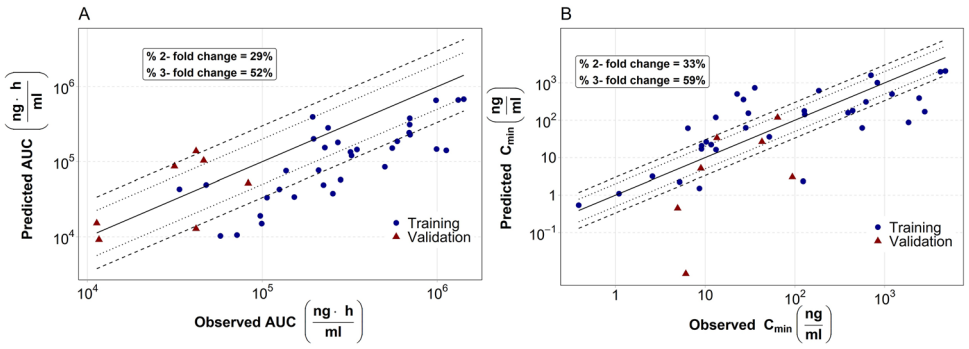

Fig. 6 Observed versus predicted area under the curve (AUC) values (**A**), and observed versus predicted minimum concentration (C_{min}) values (**B**), computed for all proteins of the training and validation sets. In both panels, percent-fold change values are also reported, representing the fraction of proteins that have a ratio between the observed and predicted values below the corresponding threshold. Dotted and dashed lines mark the 2- and threefold limits, respectively

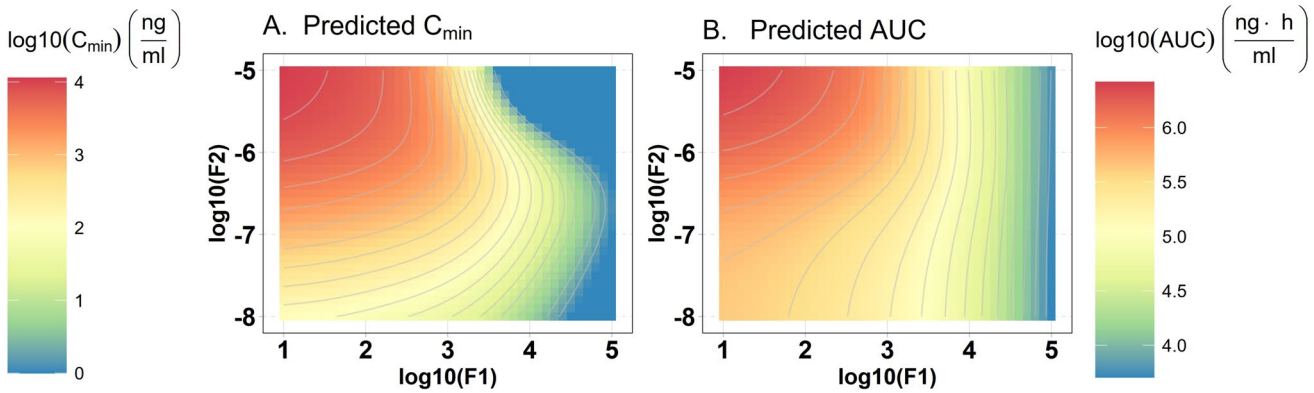

Fig. 7 Heatmaps showing the predicted values of Cmin (**A**) and AUC (**B**) as functions of F1 and F2 values from a range compatible with the proteins considered in this work. All predictions were calculated using the modified PBPK model by simulating the same initial dose of 2 mg/kg for 168 h

unmodified PBPK model with those obtained running the modified PBPK model for the training and validation sets (Fig. 5). The substantial reduction in prediction error observed in both datasets underscores the improved fidelity of the modified model. Furthermore, the comparable error distributions between training and validation sets suggest that the model exhibits strong generalizability and robust predictive capability across diverse Fc-fusion proteins.

Another metric of model accuracy is based on the comparison of the area under the curve ($AUC_{(0\text{-last})}$) of the observed and predicted PK profiles calculated using the logarithmic trapezoidal method, as reported in Fig. 6. It is customary to evaluate the prediction accuracy by computing the %-fold change, i.e., the percentage of the predicted values that are less than 2, 3, or 10 folds away from the observed values. 29%, 52% and 100% of the computed AUCs are below the three respective thresholds, confirming the fair accuracy of the predictions. As illustrated in Fig. 6A, the largest deviations from the observed AUCs are due to underestimations in the training set. This deviation from the perfect prediction is mainly due to an underestimation of the first elements of the time series, which are the most responsible for the AUC computation. For instance, in the case of proteins Fc10, Fc21, and Fc23 in Fig. 3, the model precisely fits the end of the dynamics at the expense of an underestimation of the points for $t < 100$ h.

A complementary indicator of the prediction accuracy is the minimum concentration C_{min}. Due to the monotonic trend of the PKs, it corresponds to the last sampled concentration, typically at 168 h. Figure 6B compares predicted and observed values. The percentages of the predicted values that are less than 2, 3 or 10 folds away from the observed values are 33%, 59% and 76%, respectively. With respect to AUC, larger percentages are below the 2-fold and 3-fold thresholds, but only for C_{min} there are some predictions that exceed the 10-fold threshold. These results are consistent with the chosen optimization strategy of targeting the whole PK profiles, which favors the improved prediction of metrics

tied to the overall PK, such as the AUC, instead of metrics focusing only on specific time points.

Figure 7 illustrates how different values of F1 and F2 affect the prediction of AUC and C_{min} in ranges including the values computed for the products considered in this work. To help using our results, a spreadsheet attached as Supplementary File S2 contains the values of F1, F2, AUC and C_{min} used to generate the figure, along with a macro computing F1 and F2 values from the user input and Eqs. (2, 3, 4, 5 and 6). Finally, we computed the absolute average fold error (AAFE) calculated from plasma concentrations, as in Eq. (7):

$$AAFE = 10^{\sum_{i=1}^{n} \frac{1}{n} \left| \log \frac{obs}{pred} \right|} \quad (7)$$

where n is the total number of observations in each protein's PK time series. If observed and predicted time profiles match perfectly, AAFE has a value of 1, and values below 2 are considered indications of a well-predicted time profile (32). This is the case for all our simulated time series, with a median value of 1.18 (min = 1.09; max = 1.51) across the combined training and validation sets.

Discussion

mAb-derived therapeutics can exhibit a wide range of clearance values despite sharing the same Fc domain. In this study, we focused on Fc-fusion proteins with a broad range of clearance values and sought to identify key *in silico* derived covariates. These covariates were subsequently incorporated into a physiologically based pharmacokinetic (PBPK) model through regression analysis to enhance predictive performance. While regression-based identification of covariates and PBPK modeling are both established approaches in protein-specific PK analysis (11, 33), our work introduces a distinctive combined approach: covariates derived from physicochemical properties are used not to directly predict PK parameters, but to scale parameters within a PBPK framework. This strategy allowed a better alignment between *in silico* data and *in vivo* pharmacokinetics, strengthening the pre-clinical relevance of the model.

We used the foundational PBPK model established by Shah and Betts (14) and embedded the formalism of inter-antibody variability delineated by Chen and Balthasar (34). Our workflow initially consisted of fitting the values of the factors F1 and F2, scaling the non-specific uptake rate (CL_{up}) and the lymph-to-blood flow (C_{LNLF}) in the Shah and Betts PBPK model (Fig. 2), respectively, to the available data. Subsequently, we applied a regression procedure specifically devoted to express F1 and F2 as functions of selected physicochemical properties of the proteins computed *in silico*. The final refinement consisted of an optimization of the coefficients of the formulas within the PBPK model.

The primary objective of this work was to identify proteins exhibiting anomalously rapid clearance using a priori computable *in silico* predictors, and hence constitute a tool to filter out suboptimal protein designs before they reach the wet lab. The observed improvements over the state-of-the-art PBPK mAb model, along with the accuracy of simulated pharmacokinetics in both training and validation protein sets, indicate that this objective was successfully achieved.

The development of predictive formulas was guided by the selection of a constrained set of physicochemical properties, which informed the structure of the modified PBPK model. A detailed examination of these formulas is essential to assess whether they reflect biologically meaningful mechanisms of protein metabolism or function primarily as computational constructs with limited mechanistic interpretation. Notably, several of the identified predictors correspond to molecular features previously recognized in the literature as relevant to protein clearance, both experimentally and through modeling efforts. These include the isoelectric point, the presence of positively charged surface patches, and hydrophobicity. This concordance suggests that the derived formulas serve as computationally valid approximations of complex in vivo protein dynamics and may offer mechanistic insights that warrant further biological investigation. In particular, the formula F1, which models non-specific uptake (CL_{UP}), was found to exhibit a positive correlation with charge-related descriptors, such as the isoelectric point of the full Fc-fusion protein (pIFc) and the extent of positively charged surface regions (patch_pos and patch_pos_3). These findings reinforce the relevance of electrostatic properties in governing protein disposition and support the biological plausibility of the model structure. These findings are consistent with previous observations (35, 36), and can be readily explained by enhanced electrostatic interactions. Specifically, positively charged molecules are more likely to be taken up by negatively charged tissue or cell barriers, promoting fluid-phase pinocytosis. However, the dependency of F1 on protein physicochemical properties is highly non-linear and a more comprehensive understanding of this dependency would benefit from complementary experimental studies or dedicated molecular dynamics simulations.

The factor F2 scaled lymph-to-blood return rate C_{LNLF}. While it is recognized that protein-specific features, such as the reflection factor from the interstitial space to the lymphatic system (11), can influence the protein recycling from the tissue to the blood flow (11), to our knowledge, this is the first time that C_{LNLF} is explicitly modeled as a protein-specific parameter. This choice stemmed from data-driven and computational considerations as it was the only configuration that could consistently reproduce the observed concentration–time profile across all

proteins. It is plausible that the interaction of molecules with interstitial fluid or lymph is affected by their charge and hydrophobicity (35, 37). Encouragingly, these are the same properties that appear in the formula for F2. However, the biochemical interpretation of hydrophobicity's role remains complex. While hydrophobic surface patches are generally associated with reduced solubility (38), the structure of the F2 formula includes hydrophobicity-dependent terms with both negative coefficients in the denominator and positive coefficients in direct dependence, complicating the mechanistic interpretation of underlying biological processes. Moreover, the relationship between F2 and protein properties is further modulated by the contribution of a positively charged surface patch, adding additional complexity to its interpretation.

It is important to acknowledge the sources of uncertainty within this modeling workflow and to describe the strategies employed to mitigate them. Key sources include the limited size of the available protein panel, the selection of physicochemical properties retained for regression, the computation of these property values, and the number of candidate formulas evaluated.

First, the dataset of pharmacokinetic (PK) profiles used in this study is relatively small, particularly considering the vast number of potential formulaic representations that could be explored. A limited number of proteins could lead to two main biases: (i) the capacity to predict only a specific range of clearance with good accuracy, and (ii) a model that overfits, having limited predictive power. The first concern is alleviated by the fact that the observed clearance values span several orders of magnitude, ensuring the model is trained across a broad spectrum of PK profiles. The second problem is widely common and often undermines the predictive usefulness of published models, particularly when the sample size is small. To address these concerns, we employed a computational approach based on a training/validation scheme and on a k-fold cross-validation (25) within the training group to make the results as robust as possible. Additionally, we employed a performance score that favors parsimonious formulas with fewer terms, thus avoiding overfitting and discouraging the inclusion of too many terms and covariates (see Eq. (1) in the Methods section).

The use of automated hierarchical clustering to reduce the input protein properties mitigates potential authors' bias in the identification of correlated clusters of variables. The choice of the medoid as a representative feature of each cluster serves the same purpose, although the selected properties are not necessarily those that could yield the best predictions at the end of the workflow.

Employing *in silico* physicochemical properties has the distinct advantage of enabling the computation of numerous features that could not be measured or could be measured with difficulty in an experimental setup serving as high throughput low-cost alternatives to laboratory-based assays. However, the results can be influenced by the selected computation environment (for example, by choosing the force fields in the molecular dynamics simulations). Ensuring the reliability of computed property values requires careful selection of the folded protein structure used as input and averaging across protein ensembles to enhance robustness.

The final factor to account for is the number of regression formulas to be tried. The potentially infinite number of combinations was made finite by limiting the exploration to a maximum of four terms in each regression formula and considering a maximal combination of two covariates in each sum term. Even with these constraints, and considering only a finite set of algebraic expressions (multiplication, inverse, logarithm), the total number of formulas to explore amounted to tens of millions, of which we computed the regression for about 25%. Given the model's already promising predictive performance, the likelihood of discovering a substantially better formula diminishes as exploration continues. Although in principle the exploration of formula space could continue, we considered the final model to be a good compromise between predictive accuracy and computational cost.

In conclusion, the methodological safeguards described above, together with the quality of the model fits and the biological relevance of the selected properties, provide strong support for the symbolic regression approach introduced here. The results indicate that this method can be effectively applied to real-world scenarios, even in resource-constrained settings with limited training data. It will be valuable to investigate whether the same predictors or even similar formulas could be extended to other mAb-derived therapeutics. Even if direct translatability is limited, our computational workflow could be broadly employed to reconstruct the PK of analogous or similar systems. The method outlined in this investigation is inherently flexible, adaptable to different PBPK models, as well as scaling factors. It also allows customization of formula space to explore and adapt the performance score according to the desired regression metrics, and to the number of dependent variables and covariates. To facilitate further research, we release a fully customizable reference implementation of the symbolic regression method in R, to support further research efforts (29).

Conclusion

In this work, we considered the problem of predicting aberrant PKs of mAbs-derived Fc-fusion proteins. To achieve this goal, we integrated *in silico* physicochemical properties of the proteins in a PBPK model, modifying the model with product-specific scaling factors and expressing these factors as functions of the properties themselves. This inclusion was

made possible by a custom procedure, involving the use of novel and ad hoc symbolic regressions to determine the best formulas linking the scaling factors to the protein properties, and the final optimization of the coefficients of the formulas by fitting protein-specific PK profiles. As demonstrated by different metrics, the modified PBPK model demonstrated good accuracy both in fitting the PKs of the protein in the training set, and in predicting the PKs of several products considered for validation, confirming its reliability as a tool to identify aberrant clearance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1208/s12248-026-01232-z.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement. This work was supported by Amgen Inc.

Data Availability All datasets will be made available upon request to the authors and were published in Bryniarski *et al.* (15).

Declarations

Disclosures K.D.C., A.C., I.F., and V.A.T. were full-time employees and shareholders of Amgen Inc. at the time of this work. D.T., A.P., R.V., and L.M. were contracted by Amgen Inc. while this research was conducted.

References

1. Kinch MS, Kraft Z, T S. Monoclonal antibodies: trends in therapeutic success and commercial focus. Drug Discov Today. 2023;28(1):103415.
2. Ramdani Y, Lamamy J, Watier H, Gouilleux-Gruart V. Monoclonal antibody engineering and design to modulate FcRn activities: a comprehensive review. Int J Mol Sci. 2022;23(17):9604.
3. Czajkowsky DM, Hu J, Shao Z, Pleass RJ. Fc-fusion proteins: new developments and future perspectives. EMBO Mol Med. 2012;4:1015–28.
4. Duivelshof BL, Murisier A, Camperi J, Fekete S, Beck A, Guillarme D, et al. Therapeutic Fc-fusion proteins: current analytical strategies. J Sep Sci. 2021;44(1):35–62.
5. Liu L. Pharmacokinetics of monoclonal antibodies and Fc-fusion proteins. Protein Cell. 2017;9(1):15–32.
6. Han JJ. FDA modernization act 2.0 allows for alternatives to animal testing. Artif Organs. 2023;47(3):449–50.
7. Boswell CA, Tesar DB, Mukhyala K, Theil FP, Fielder PJ, Khawli LA. Effects of charge on antibody tissue distribution and pharmacokinetics. Bioconjugate chemistry. 2010;21(12):2153–2163.
8. Datta-Mannan A. Balancing charge in the complementarity-determining regions of humanized mAbs without affecting pI reduces non-specific binding and improves the pharmacokinetics. MAbs. 2015;7(3):483–93.
9. Liu S. Effect of variable domain charge on in vitro and in vivo disposition of monoclonal antibodies. MAbs. 2021;13(1):1993769.
10. Jain T, Sun T, Hall A, Houston NR, Nett JH, Sharkey B, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci U S A. 2017;114(5):944–9.
11. Liu S, Humphreys S, Cook K, Conner K, Correia A, Jacobitz A, et al. Utility of physiologically based pharmacokinetic modeling to predict inter-antibody variability in monoclonal antibody pharmacokinetics in mice. MAbs. 2023;15(1):2263926.
12. Gruber A, Führer F, Menz S, Diedam H, Göller AH, Schneckener S. Prediction of human pharmacokinetics from chemical structure: combining mechanistic modeling with machine learning. J Pharm Sci. 2024;113(1):55–63.
13. Conner KP,DSC,TVA,&RDA. The biodistribution of therapeutic proteins: Mechanism, implications for pharmacokinetics, and methods of evaluation. Pharmacology & therapeutics. 2020;212:107574.
14. Shah DK, Betts AM. Towards a platform PBPK model to characterize the plasma and tissue disposition of monoclonal antibodies in preclinical species and human. J Pharmacokinet Pharmacodyn. 2012;39(1):67–86.
15. Bryniarski MA, Wang S, Chen A, Coventry B, Korkmaz EN, Haque Tuhin MT, et al. Nonspecific cellular interactions are a key determinant in the disposition of fc-fused proteins. Mol Pharm. 2026;23(2):59–882.
16. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:85–589.
17. Thorsteinson N, Gunn J, Kelly K, Long W, Labute P. Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. MAbs. 2021;13(1):1981805.
18. Labute P. Lowmodemd—implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. J Chem Inf Model. 2010;50(5):792–800.
19. Pardo S. Regression and Model Fitting with Collinearity. SpringerLink; 2020.
20. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw. 2014;61:1–36.
21. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963. https://doi.org/10.2307/2282967.
22. Augusto DA, Barbosa HJ. Symbolic regression via genetic programming. In: Proceedings. Vol. 1. Sixth Brazilian symposium on neural networks. IEEE;2000:173–178.
23. Valipour M. SymbolicGPT: A Generative Transformer Model for Symbolic Regression. ArXiv. 2021.
24. Kim S, Lu PY, Mukherjee S, Gilbert M, Jing L, Ceperic V, Soljacic M. Integration of neural network-based symbolic regression in deep learning for scientific discovery. IEEE Trans Neural Netw Learn Syst. 2020;32(9):4166–4177.
25. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer; 2009, pp. 1–758.
26. Myrvold W, Ruskey F. Ranking and unranking permutations in linear time. Inf Process Lett. 2001. https://doi.org/10.1016/S0020-0190(01)00141-7.
27. Scrucca L. GA: a package for genetic algorithms in R. J Stat Softw. 2013;53:1–37.

28. Michie D. “Memo” functions and machine learning. Nature. 1968. https://doi.org/10.1038/218306c0.
29. Tomasoni D. [Online].; 2024. Available from: https://doi.org/10.5281/zenodo.12904322.
30. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(1):86–97.
31. Nelder JA, Mead R. A simplex method for function minimization. Comput J. 1965;7(4):308–313.
32. Saeheng T, Na-Bangchang K, Karbwang J. Utility of physiologically based pharmacokinetic (PBPK) modeling in oncology drug development and its accuracy: a systematic review. Eur J Clin Pharmacol. 2018;74:1365–76.
33. Zou P. Predicting human bioavailability of subcutaneously administered fusion proteins and monoclonal antibodies using human intravenous clearance or antibody isoelectric point. AAPS J. 2023;25:31.
34. Chen Y, Balthasar JP. Evaluation of a catenary PBPK model for predicting the in vivo disposition of mAbs engineered for high-affinity binding to FcRn. AAPS J. 2012;14(4):850–859.
35. Hu S, Datta-Mannan A, D’Argenio D. Monoclonal antibody pharmacokinetics in cynomolgus monkeys following subcutaneous administration: physiologically based model predictions from physiochemical properties. AAPS J. 2023;25:5.
36. Schoch A, Kettenberger H, Mundigl O, Winter G, Engert J, Heinrich J, et al. Charge-mediated influence of the antibody variable domain on FcRn-dependent pharmacokinetics. Proc Natl Acad Sci U S A. 2015;112(19):5997–6002.
37. Sharmaa V, Patapoffa T, Kabakoffa B, Paia S, Hilarioa E, Zhang B, et al. In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. PNAS. 2014;111(52):18601–6.
38. van Gils J, Gogishvili D, van Eck J, Bouwmeester R, van Dijk E, Abeln S. How sticky are our proteins? Quantifying hydrophobicity of the human proteome. Bioinform Adv. 2022. https://doi.org/10.1093/bioadv/vbac002.
39. Larson S. The shrinkage of the coefficient of multiple correlation. J Educ Psychol. 1931;22(1):45–55.