MABS
2025, VOL. 17, NO. 1, 2474521
https://doi.org/10.1080/19420862.2025.2474521

Taylor & Francis
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

PROPERMAB: an integrative framework for *in silico* prediction of antibody developability using machine learning

Bian Li[a], Shukun Luo[b], Wenhua Wang[b], Jiahui Xu[b], Dingjiang Liu[b], Mohammed Shameem[b], John Mattila[c], Matthew C. Franklin[a], Peter G. Hawkins[d*], and Gurinder S. Atwal[d#]

[a]Therapeutic Proteins, Regeneron Pharmaceuticals, Inc, Tarrytown, NY, USA; [b]Formulation Development, Regeneron Pharmaceuticals, Inc, Tarrytown, NY, USA; [c]Preclinical Manufacturing and Process Development, Regeneron Pharmaceuticals, Inc, Tarrytown, NY, USA; [d]Molecular Profiling and Data Science, Regeneron Pharmaceuticals, Inc, Tarrytown, NY, USA

ABSTRACT

Selection of lead therapeutic molecules is often driven predominantly by pharmacological efficacy and safety. Candidate developability, such as biophysical properties that affect the formulation of the molecule into a product, is usually evaluated only toward the end of the drug development pipeline. The ability to evaluate developability properties early in the process of antibody therapeutic development could accelerate the timeline from discovery to clinic and save considerable resources. *In silico* predictive approaches, such as machine learning models, which map molecular features to predictions of developability properties could offer a cost-effective and high-throughput alternative to experiments for antibody developability assessment. We developed a computational framework, PROPERMAB (PROPERties of Monoclonal AntiBodies), for large-scale and efficient *in silico* prediction of developability properties for monoclonal antibodies, using custom molecular features and machine learning modeling. We demonstrate the power of PROPERMAB by using it to develop models to predict antibody hydrophobic interaction chromatography retention time and high-concentration viscosity. We further show that structure-derived features can be rapidly and accurately predicted directly from sequences by pre-training simple models for molecular features, thus providing the ability to scale these approaches to repertoire-scale sequence datasets.

ARTICLE HISTORY
Received 5 November 2024
Revised 25 February 2025
Accepted 26 February 2025

KEYWORDS
Antibody developability; feature engineering; *in silico* prediction; machine learning

Introduction

Monoclonal antibody (mAb)-based biologics have become a major therapeutic modality in recent years.[1,2] As of December 2023, more than 100 antibody therapeutics have been marketed and the number of antibody therapeutics in late-stage clinical studies has surpassed 130.[3] For several disease areas such as oncology,[4,5] inflammation, and infectious disease,[6] antibodies have become the predominant treatment options.[1]

Despite the continuous advances in antibody discovery and development technologies, bringing an antibody from discovery to a marketed product remains a costly process with a high attrition rate.[7] While termination of clinical development can result from various efficacy and safety issues, in the early development stage poor developability properties pose unique challenges in product formulation and Chemistry, Manufacturing, and Control (CMC) process development.[7–9] A successful antibody drug development program relies on the selection of a lead candidate that is developable, with preferred properties in a variety of developability attributes such as expression titer, purification yield, solubility, viscosity, stability, and administration compatibility. Poor developability properties may add considerable cost to development and slow the timeline from discovery to clinic. Some developability issues such as high viscosity can be mitigated by optimization of excipients,[10,11] albeit often resulting in a deviation from platform formulation and manufacturing process. Other developability issues such as increased hydrophobicity or poor stability may require extensive optimization of purification process or even engineering of the sequence.[12] Given the considerable time and cost associated with off-platform manufacturing process and formulation or failure of development, it is essential to prioritize sequences that are less likely to pose developability issues when candidates are selected.[7,8]

In early-stage screening, candidate molecules are typically selected from a library of well over a thousand antibodies. Experimental assessment of the developability properties of such a library is challenging if not infeasible. This is in part due to the requirement for sufficient purified materials for each antibody and/or the development of high-throughput assays.[13,14] As an alternative, computational methodologies for developability assessment are gaining increased attention.[15] For example, several computational strategies have recently been introduced for predicting antibody solubility,[16,17] aggregation,[18–20] thermostability,[21] hydrophobic interaction chromatography (HIC) retention time,[20,22] pharmacokinetics,[20,23,24] and high-concentration viscosity.[20,23,25–28]

CONTACT Bian Li bian.li@regeneron.com Therapeutic Proteins, Regeneron Pharmaceuticals, Inc, 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA; Peter G. Hawkins peter.hawkins@databricks.com; Gurinder S. Atwal matwal@flagshippioneering.com

[*]Current address: Databricks, 5 Bryant Park, 8th floor, New York, NY 10018.
[#]Current address: Flagship Pioneering.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2025.2474521

Historically, *in silico* developability assessment strategies have typically been heuristics that are based on individual molecular features designed to associate with a specific developability attribute[29,30] or simple linear statistical or empirical models that consider only a handful of features.[31,32] More recently, machine-learning (ML) models have been increasingly used for *in silico* developability assessment.[14,16,21–23,26,27] In addition to larger and higher-quality experimentally derived training data, having informative features as input is critical to the predictive performance of ML models. Several commercial software packages are available and can calculate a large set of input features. However, users are typically limited to a set of predefined features and implementing novel, user-designed features can be challenging in part due to the lack of source code access. Moreover, these software packages are typically not integrated with the Python ML ecosystem, thus including them into automated, Python-based ML model development workflows can be challenging.

Here, thanks to recent advances in 3D structure prediction and sequence modeling for proteins[33–35] and antibodies,[36,37] we developed a versatile and integrative computational tool called PROPERMAB to enable *in silico* developability assessment of mAbs using machine learning. This is accomplished by PROPERMAB's ability to efficiently calculate a diverse set of features, ease of engineering new features, and through its seamless integration with the Python ML ecosystem. We show the practical utility of PROPERMAB by applying it to predict mAb HIC retention time and high-concentration viscosity. We also demonstrate a computationally efficient and practically effective approach for applying PROPERMAB to repertoire-scale sequence datasets.

Results

Design and overview of PROPERMAB

We developed PROPERMAB to facilitate *in silico* prediction of mAb developability properties with the only assumption that mAb sequences are available at the point of assessment. The design of PROPERMAB was guided by the preference for integrating features at both sequence and structure levels, ease of engineering new features, and the need for robust model validation, especially in the data-limited, supervised learning regime such as antibody developability prediction. At a high level, given the heavy and light-chain sequences of a mAb, one can use PROPERMAB to calculate a predefined set of sequence-based features (the Sequence Featurizer component in Figure 1, and Supplementary Table S1), extract sequence embeddings from pretrained protein language models (PLMs)[34,37] (not shown in Figure 1), employ a deep learning-based protein 3D structure prediction method to predict the 3D structure for the Fv domain[36] (the Structure Predictor component in Figure 1), and calculate a predefined set of structure-based features from the predicted structures (the Structure Featurizer component in Figure 1, and Supplementary Table S2). Once features have been calculated both supervised and unsupervised learning can be pursued depending on the availability of experimentally measured developability data and the questions being asked. To facilitate the development of supervised models, we designed a ModelTrainer class that provides a unified interface for model training and validation (the Model Trainer component in Figure 1). The ModelTrainer class interfaces with the scikit-learn package[38] and encapsulates robust cross-validation

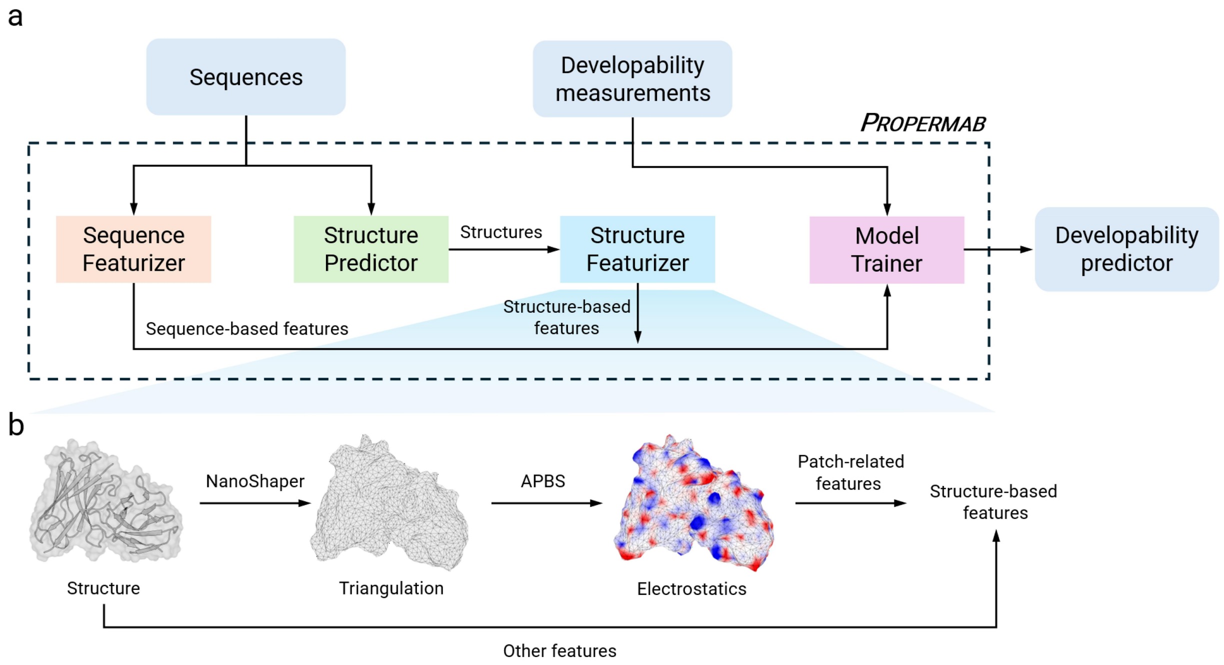


Figure 1. A schematic of PROPERMAB highlighting its main components. a) The main components of PROPERMAB were designed to facilitate feature calculation and supervised ML model development. Given antibody sequences, one uses the sequence featurizer component to calculate sequence-based features. To calculate structure-based features, one first uses the structure predictor component to predict 3D structures, which are then used as input to the structure featurizer component for feature calculation. Both the sequence and structure featurizers currently offer a diverse set of features and can be easily extended with new, custom-designed features. Once features have been calculated, the training and validation of ML models are done through the model trainer component. b) A zoomed-in view of the key steps and components of the structure featurizer.

strategies for model parameter fitting, hyper-parameter tuning, and performance estimation.

PROPERMAB was implemented purely in Python and made open source (for noncommercial use). As such, it offers algorithmic transparency, ease of engineering new features, and is seamlessly integrated with other Python-based, open-source ecosystems for scientific computing such as machine learning and bioinformatics. In the following sections, we first demonstrate the ease of engineering new features in PROPERMAB by showcasing the implementation of a set of novel structure-based features that, to our knowledge, have not previously been used in mAb developability prediction. We then discuss examples of using PROPERMAB to develop ML models for predicting mAb HIC retention time and high-concentration viscosity, two important properties considered during mAb therapeutic product and process development. Finally, because predicting 3D structures and calculating structure-based features can be computationally costly for large, multi-repertoire-scale datasets, we demonstrate that one can train simple regularized linear regression models to accurately predict structure-based features directly from sequence. Using predicted features as input results in only a slight drop in performance compared with using features calculated from structures, suggesting a cost-effective approach for large datasets.

PROPERMAB implements a diverse set of molecular features and facilitates engineering of new features

PROPERMAB currently implements a diverse set of molecular features that are either derived from sequences (9 features, Supplementary Table S1) or structures (26 features, Supplementary Table S2). These features characterize the biophysical attributes of antibodies, such as electrostatic charge distribution, hydrophobicity, and solvent exposure, either based on amino acid sequences or the 3D structures of the Fv domain. For several charge and hydrophobicity-related features, PROPERMAB also offers versions specific to or near the complementarity-determining regions (CDRs) because they play the central role in antibody–antigen interaction. These features were largely inspired by current understanding of the molecular and biophysical mechanisms contributing to poor solution and colloidal properties and the fact that some of them have been used in the existing literature.[14,24,39–41] For example, the utility of surface patches (Figure 1, Methods, and Supplementary Table S2) in antibody developability assessment has been demonstrated in several recent studies.[20,24,40,41]

Of the current set of features computed by PROPERMAB, several of the sequence-derived, charge- or hydrophobicity-related features have been previously reported and used for *in silico* developability assessment.[22,27] However, PROPERMAB allows easy customization of some of the previously reported features for specific antibody regions. For example, we customized the net_charge feature for the CDRs (net_charge_cdr) and solvent-exposed residues (exposed_net_charge). Because PROPERMAB implements a triangle mesh-based strategy to compute surface patches, the vertices of the triangle mesh can be assigned with various biophysical properties and can also be mapped to residues (Figure 1 and Methods). Such information can be leveraged to create customized features such as number of patches and largest patch area of various physicochemical nature (i.e., positively charged, negatively charged, hydrophobic) and for specific antibody regions. More importantly, PROPERMAB also enables the design of novel features to leverage expert domain knowledge on the developability property of interest. For example, hypothesizing that the spatial pattern of biophysical features affects inter-molecular interactions, we implemented a variant of the Ripley's K statistic[42,43] for charged and aromatic residues (Figure 2 and Supplementary Table S2) as features for developability assessment. The Ripley's K is a statistic that characterizes spatial correlation of point patterns and is defined as the fraction of point pairs for which the distance is less than a distance cutoff (Figures 2(a,b)). Our implementations of the Ripley's K statistics characterize the clustering or dispersion of charge and aromatic features over the surface of antibody Fv domain at a given distance threshold (Figure 2(c)). As discussed in the next two sections, some of the Ripley's K features are proved to be informative in predicting antibody HIC retention time and viscosity.

Application of PROPERMAB to predict HIC retention time

We developed PROPERMAB as a general framework for feature engineering and ML model development for mAb developability prediction. As examples of use cases, we first demonstrate PROPERMAB's utility by applying it to predict HIC retention time (HIC RT) and in the next section we apply it to predict high-concentration viscosity.

When screening for candidate molecules, it is important to consider a molecule's hydrophobicity profile because increased hydrophobicity of mAbs often contributes to higher likelihood of precipitation and aggregation,[41,44] reduced purification yield or off-platform purification process. HIC is a standard biophysical method commonly used in protein purification[45] and HIC RT is characteristic of individual molecule's surface hydrophobicity with longer RTs correlating with higher degree of hydrophobicity.[46] To demonstrate the utility of PROPERMAB, we computed all 35 features for 135 mAbs for which the HIC RTs have been consistently measured[47] and first analyzed the correlations of individual features with HIC RT (Figure 3(a)). To account for the effect of 3D conformation on feature values (Supplementary Figure S1 and Supplementary Table S3), we computed structure-based features across five predicted structures and took the averages. Our analysis shows that about two-thirds of the features (23 out of 35) are significantly correlated with HIC RT (Figure 3(a)), among which hyd_patch_area_cdr (total area of hydrophobic patches near CDRs, Methods) has the highest correlation, suggesting that mAbs likely bind the stationary phase of the HIC column through hydrophobic patches on their CDRs. In addition to hyd_patch_area_cdr, our analysis shows that the aromatic_asa feature (aromatic surface area) also has a strong correlation with HIC RT

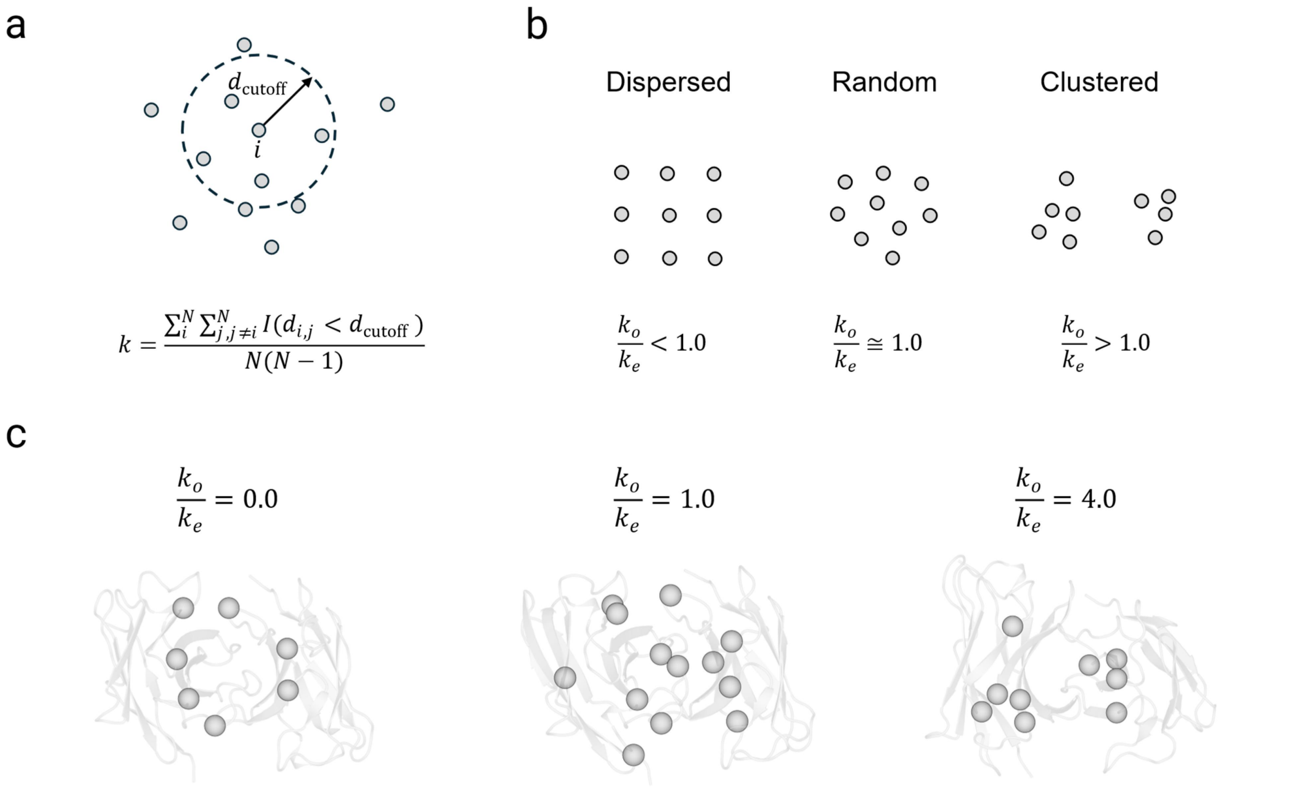


Figure 2. Implementation of a variant of the Ripley's K function as molecular features. a) An illustration of the computation of the K function at contact distance cutoff d_{cutoff}. In the formula, N is the total number of points, $d_{i,j}$ is the Euclidean distance between point i and point j, I is the indicator function that evaluates to 1 when $d_{i,j}$ is less than d_{cutoff} and 0 otherwise. Thus, k is the fraction of point pairs for which the distance is less than d_{cutoff}. b) Schematics showing three different spatial patterns of point clouds. Here, k_o is the Ripley's K based on the observed pattern, k_e is the expected Ripley's K when the points are randomly scattered. Note that the d_{cutoff} in this panel is much smaller than the d_{cutoff} in panel a). c) Examples of the Ripley's K spatial statistic feature for surface aromatic residues. Here, only the Cα atoms of solvent-exposed aromatic residues are shown. Distances are measured between pairs of Cα atoms and d_{cutoff} is set to 6 Å.

(Figures 3(a,c)). Aromatic residues have previously been found to be enriched in aggregation prone regions in mAbs.[49] We also note that two novel, Ripley's K derived features, neg_ripley_k (Ripley's K ratio for negatively charged residues) and aromatic_ripley_k (Ripley's K ratio for aromatic residues), are also found to be significantly correlated with HIC RT (Figures 3(a,d)). While the correlations are only moderate, it nevertheless demonstrates that one can engineer rather sophisticated and predictive features using our framework.

Encouraged by the correlations of features with HIC RT, we then trained an ElasticNet model using all molecular features currently available in PROPERMAB as input to predict HIC RT. We used a nested cross-validation strategy to estimate the performance of the model (Methods). Our nested leave-one-out cross-validation (LOOCV) evaluation shows that the ElasticNet regression model predicts HIC RT with a Pearson's $r = 0.71$ and a Spearman's $\rho = 0.75$, substantially better than any of the individual features (Figure 3(e)). We note that due to the small size of the dataset, we opted not to do explicit feature selection. However, after inspecting the model coefficient associated with each feature, we found that the set of features with a non-zero coefficient in the ElasticNet model is highly concordant with the set of features that are significantly correlated with HIC RT (Figure 3(a)).

Application of PROPERMAB to predict high-concentration viscosity

Viscosity is a key developability attribute that can significantly affect both manufacturing and subcutaneous injection. mAb solutions with high viscosities add challenges to manufacturing, delivery, and administration.[50] To demonstrate the utility of PROPERMAB in predicting viscosity, we applied it to an internal viscosity dataset consisting of 58 unique and diverse IgG4 mAbs (on average the sequences of VH domains in this dataset differ by 42 residues from each other and the sequences of VL domains differ by 40 residues) and two previously published datasets of the IgG1 isotype.[25,51] The viscosities of mAbs in these datasets have been experimentally determined at a concentration of 150 mg/mL. To our knowledge, our IgG4 dataset is the largest dataset of measured viscosities of the IgG4 isotype reported at the time of writing. As with HIC RT, our analysis shows that several features are strongly correlated with viscosity in the two previously published datasets (Supplementary Figure S2). However, no single feature is strongly correlated with viscosity in our internal IgG4 dataset (Figure 4(a)). This discrepancy is likely due to the fact that the two previously published datasets are of the IgG1 isotype and that molecules in one of the datasets, PDGF38, were derived from the same parent molecule and designed to optimize surface electrostatic properties for better viscosity profile.[51] Nevertheless,

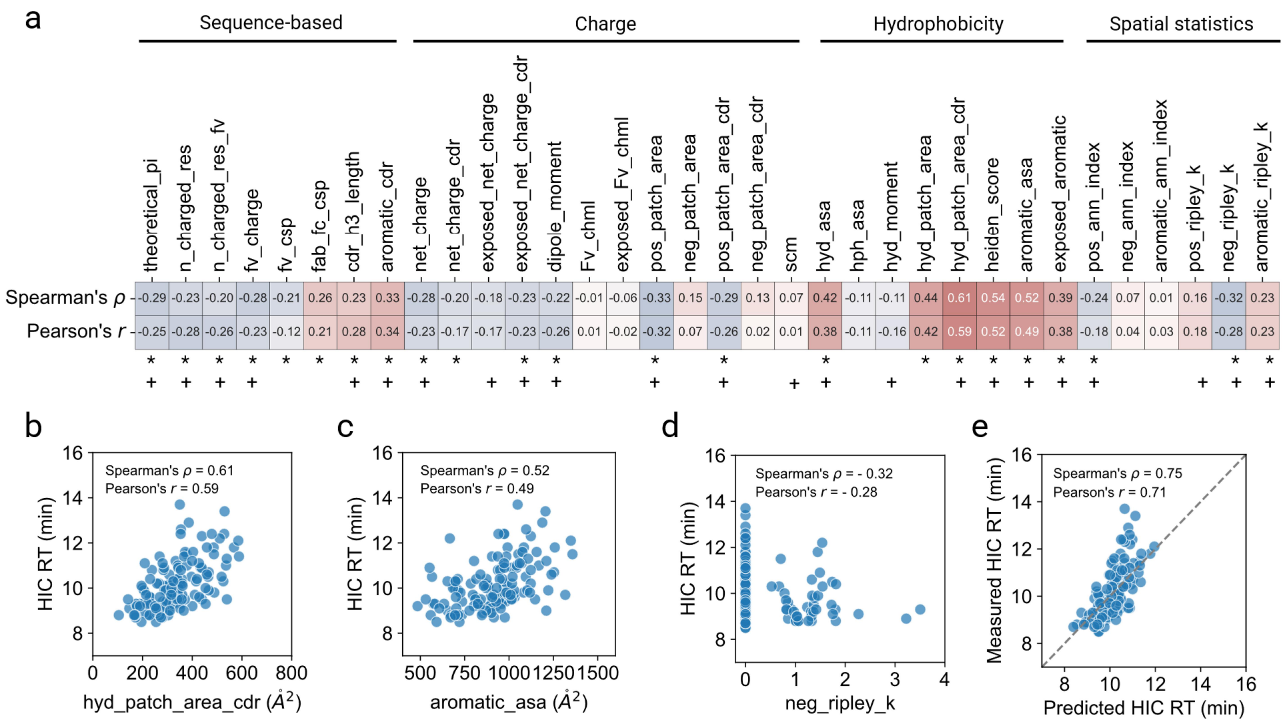


Figure 3. Applying PROPERMAB to predict mAb HIC RT. a) Spearman's ρ and Pearson's *r* correlation coefficients of each feature with the HIC RTs of 135 mAbs from Jain et al.[47] We excluded the two mAbs for which the HIC RTs were not determined but arbitrarily set to 25 min. The fc_charge feature is omitted because these molecules have identical constant region sequences,[47] which resulted in identical values of the fc_charge feature across all 135 mAbs. Features with a statistically significant ($p < 0.05$, adjusted via the Benjamini-Hochberg procedure)[48] Spearman's ρ are indicated with a * symbol. b, c, d) Scatter plots of HIC RT vs. a few examples of molecular features. The hyd_patch_area_cdr feature (total hydrophobic patch area near CDRs) has the highest correlation with HIC RT, which is followed by the heiden_score and aromatic_asa features (total solvent accessible surface area contributed by aromatic residues). The spatial statistic neg_ripley_k, which was designed to characterize the clustering or dispersion of negatively charged residues on antibody surface, has a moderate but significant negative correlation (Spearman's ρ = −0.32, adjusted $p < 0.05$) with HIC RT. A neg_ripley_k of 0 indicates that there was no pair of negatively charged residues that are less than 6 Å from each other. e) Leave-one-out cross-validation results of an ElasticNet regressor trained to predict mAb HIC RT using all the features. Features with a non-zero coefficient in the ElasticNet model are indicated with a + symbol in panel a).

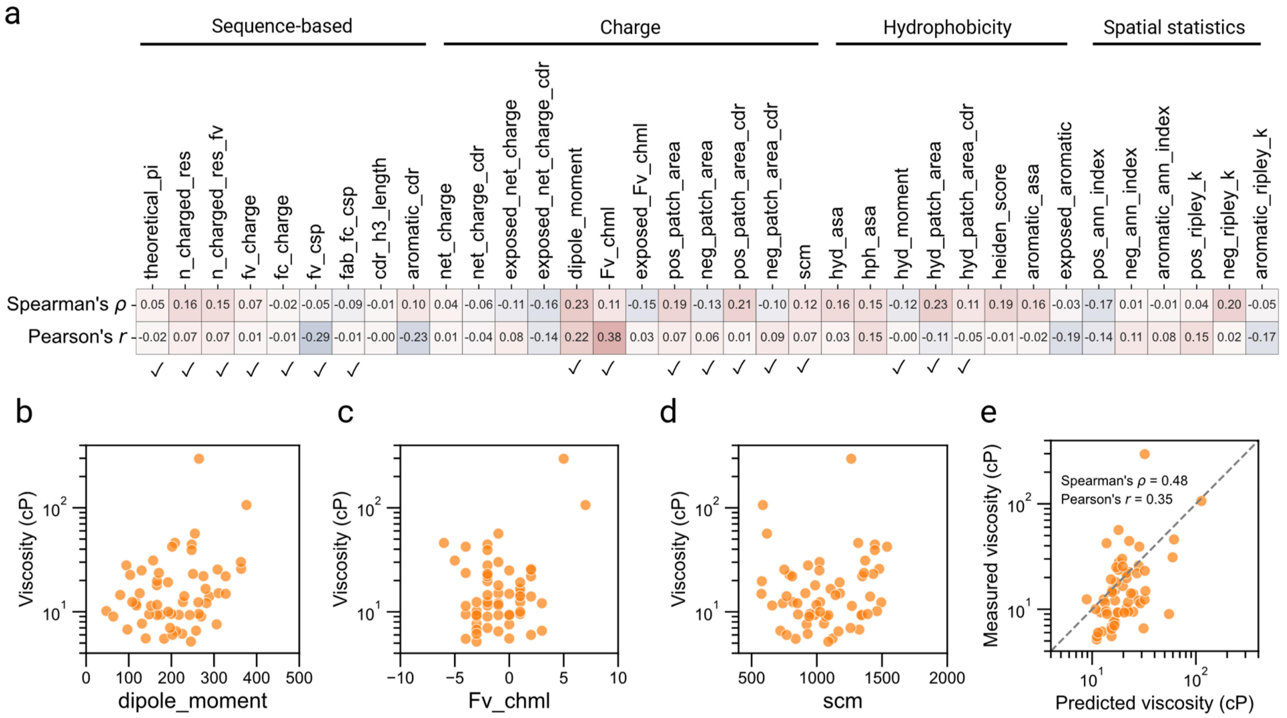


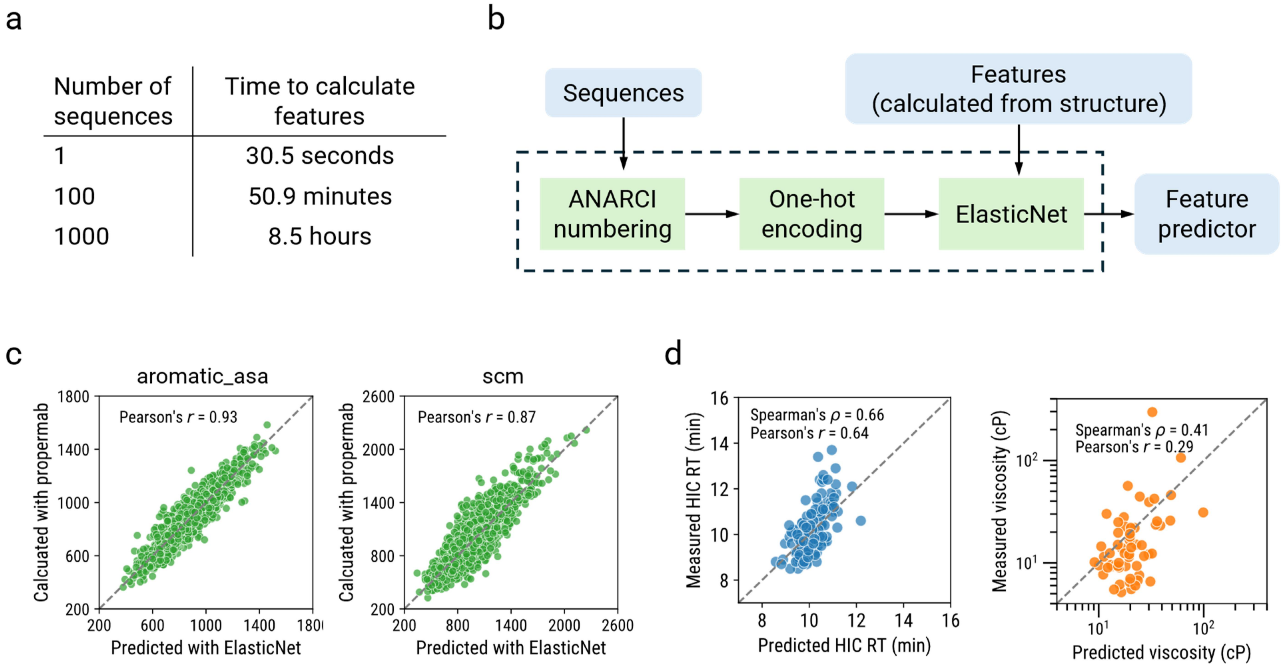
Figure 4. Applying PROPERMAB to predict high-concentration viscosity of IgG4 mAbs. a) Spearman's ρ and Pearson's *r* correlation coefficients of each feature with IgG4 mAb viscosity. While there was no single feature that is strongly correlated with viscosity, several charge asymmetry related features such as dipole_moment and Fv_chml demonstrate at least weak correlation with viscosity. b, c) Scatter plot of viscosity vs. dipole_moment and Fv_chml respectively, with both having at least some weak correlation with viscosity. d) Scatter plot of viscosity vs. the scm feature, showing almost no correlation between the two. e) Leave-one-out cross-validation results of a random forest regressor trained to predict viscosity using a set of features manually selected based on in-house expertise (indicated with a ✓ symbol in panel a).

several features characterizing charge asymmetry across the VH and VL domains such as dipole_moment and Fv_chml were found to be at least weakly correlated with viscosity in our IgG4 dataset (Figures 4(a,b,c)). It has previously been shown in a much smaller dataset (14 IgG4 mAbs) that the spatial charge map (SCM) score[29] can correctly classify IgG4 mAbs into low or high viscosity with ~80% accuracy.[52] This finding is not supported by our analysis of a much larger dataset because we find almost no correlation between the SCM score and viscosity (Spearman's $\rho = 0.12$ and a Pearson's $r = 0.07$) (Figures 4(a,d)). While the molecular origins of viscosity behaviors of mAb solutions are believed to be quite complex,[53] our findings with a larger dataset provide strong support for the role that charge asymmetry plays in IgG4 mAb viscosity.

We next sought to develop an ML model to predict mAb viscosity. Based on the understanding that hydrophobic and electrostatic interactions are two major contributing factors of high viscosity, we selected a set of hydrophobicity and charge-related features as model input. We note that the selection of features was done prior to examining their correlations with viscosity to avoid information leakage[54] and no algorithmic feature selection was done due to the limited training data available. As an example, we trained a random forest model because of its ability in modeling highly non-linear relationships. We followed the same model building and nested cross-validation evaluation procedure as in HIC RT prediction (Methods). As shown in Figure 4(e), the random forest model combining multiple features of diverse biophysical nature as input achieved a Spearman's $\rho = 0.48$ and a Pearson's $r = 0.35$. Together with the example on HIC RT prediction, our results demonstrate a broad applicability of the PROPERMAB framework to *in silico* prediction of mAb developability properties.

Computational efficiency enables large-scale assessment

Calculating structure-derived features given mAb Fv sequences using PROPERMAB takes about 30 seconds on average on an AWS instance with 8 CPU cores and 30GB of RAM (Figure 5(a)). Our calculation suggests that the time and cost associated with calculating the features for a typical secondary screening dataset (~1000 sequences) is fast and cost-effective (8.5 hours with \$0.384/hour). While the computing time scales only linearly with the number of mAb Fv sequences (Supplementary Figure S3), the time and computational cost could be prohibitive if one were to apply PROPERMAB to large repertoire sequence projects where the scale of sequences can reach millions. We reasoned that if the structure-derived features can be predicted directly from sequence using simple lightweight ML models, then a large fraction of the computing cost in structure prediction and feature calculation can be saved (Figure 5(b)). As examples, we trained ElasticNet models to predict all structure-derived features from sequences alone. We randomly sampled 12,000 unique mAbs from the OAS database[55] and divided them into a training set of 10,000 mAbs and a test set with 2,000 mAbs. For each of the mAbs, we calculated the structure-based features using PROPERMAB and one-hot encoded the sequences after aligning them



Number of sequences	Time to calculate features
1	30.5 seconds
100	50.9 minutes
1000	8.5 hours

Figure 5. PROPERMAB enables efficient large-scale feature prediction from sequences. a) Time cost of calculating structure-based features using PROPERMAB. We ran and timed PROPERMAB on 1000 paired sequences randomly sampled from the OAS database.[55] The time for one sequence is the average across 1000 sequences. The time for 100 sequences is the average across a 1000 bootstrap samples. Each bootstrap sample was obtained by randomly sampling 100 individual sequences from the 1000-sequence run. b) A schematic showing our strategy to train ElasticNet models to predict structure-based features from sequences. c) Comparison of feature values calculated based on structures with values predicted from sequences using trained ElasticNet models for aromatic_asa and scm, respectively. The plots are based on a test set of 2000 paired sequences randomly sampled from the OAS database.[55] d) Comparison of measured values of HIC RT and viscosity with values predicted using models trained with predicted molecular features as input.

according to the IMGT numbering scheme[56,57] (Figure 5(b)). (Calculating features for a set of 12,000 mAbs would incur one-time cost.) Alignment of sequences to the IMGT numbering scheme[56] enables one-hot encoding with fixed size and biologically meaningful method of padding. Our results show that the structure-derived features can be predicted from sequences alone with high accuracies, with a median Pearson's *r* of 0.87 (Figure 5(c) and Supplementary Figure S4). Encouraged by these results, we applied the trained ElasticNet models to predict the structure-based features for the HIC RT and viscosity datasets. While models trained with the predicted features as input experience a slight drop in performance (Figure 5(d)) compared to models trained with features calculated directly from structures (Figures 3(e) and 4(e)), predicting features from sequence is considerably faster than calculating features from structures. In fact, on a MacBook Pro laptop with the M1 Pro processor and 16GB of RAM, it took less than 2 minutes to predict all 26 structure-based features for >140k paired sequences from the OAS database once the sequences have been one-hot encoded. On the same computing platform, numbering the >140k sequences according to the IMGT scheme and one-hot encoding them took less than 3 hours. Thus, given the accuracy and efficiency of sequence-based prediction, it is feasible to apply PROPERMAB to large repertoire sequence datasets.

Discussion

Screening candidates for lead molecules with better developability properties early in the discovery and development of antibody therapeutics has increasingly been recognized as important for success.[7,11,13,15] While multiple biophysical assays have been developed over the years,[47] ML models integrating computational molecular features and trained on experimental data could offer a cost-effective approach.[13] We developed PROPERMAB as an integrative framework with the capability of calculating diverse molecular features and enabling downstream ML model development using these features.

While PROPERMAB can be applied to a broad spectrum of developability properties, we highlighted its applicability with two use cases. In the case of HIC RT prediction, we showed that several features calculated by PROPERMAB demonstrated moderate to strong correlation with HIC RT. When an ElasticNet model was trained, it resulted in much higher performance even though about one-third of the features were eliminated (i.e., zero model coefficient). The elimination of several features is not unexpected given that some of them are not significantly correlated with HIC RT and that there is some redundancy in the feature space (Supplementary Figure S5). We chose the ElasticNet model primarily because it is easier to interpret, but results from another study indicate that HIC RT is likely related to charge and hydrophobicity in a non-linear manner.[58] This suggests that combining charge and hydrophobicity-related features into a non-linear model may further improve prediction accuracy, although the resulting model may become less interpretable.

In the case of predicting high-concentration viscosity, we showed that several individual features are highly correlated with viscosity in two previously published datasets of the IgG1 isotype, but no features had strong correlation with viscosity in our IgG4 dataset. Compared to IgG1 mAbs, there is much less viscosity data of IgG4 mAbs. Our finding based on a relatively large internal IgG4 dataset suggests that the mechanisms underlying high viscosity of IgG4 mAbs are likely to be more complex than IgG1 mAbs and that identifying a single, strongly predictive feature requires better understanding of the underlying biophysical mechanisms. The difference in predictive power of individual features between IgG1 and IgG4 mAbs also suggests that performances for mAbs of other types or isotypes may vary; this warrants further investigations. Nevertheless, when multiple features were integrated into an ML model, prediction of IgG4 mAb viscosity is improved. We note that in the current work the predictive power of individual features is evaluated based on their linear correlations, which can fail to quantify associations that are nonlinear.[59] A possible future development in feature evaluation is to employ an information-theoretic formalism in contrast to the traditional correlation-based approaches. For example, by quantifying the mutual information, a self-equitable measure of association, between features and the biophysical properties, one could identify features that maximize information gain and thereby improve prediction.[59]

Recently, there has been interest in applying *in silico* developability assessment tools to large repertoire sequence datasets that may contain $\sim 10^{10}$ sequences. While structure-derived features are critical for antibody developability prediction,[60] calculating the structure-derived features for each molecule at this scale is computationally prohibitive. Here, we showed that structure-derived features as calculated using PROPERMAB can be predicted with high accuracy from sequence alone with relatively simple models. Previously, deep learning surrogate models were trained to predict mAb spatial properties such as the SCM score with target features in the training sets calculated as ensemble averages across 10-ns long molecular dynamics (MD) trajectories.[61,62] While computationally more expensive structure modeling such as MD simulations may generate important mechanistic insights, recent results on its benefit to ML-based developability prediction are not yet consistent.[20,22] In this work, we demonstrated a computationally much more efficient and practically effective alternative to predicting structure-based features from sequences only. We trained simpler ElasticNet models to predict features calculated by PROPERMAB and avoided the use of expensive MD simulations. Importantly, we showed that ML models using ElasticNet predicted features as input resulted in only a slight drop in performance. The efficiency and effectiveness of our method suggest that it can be a viable approach for *in silico* developability prediction of repertoire-scale datasets. However, we also note that there is a substantial variation in prediction accuracy across the structure-derived features. Understanding why some features are less accurately predicted than others would provide important insights into the limitations of predicting structure features from sequences.

Additionally, it was recently shown that there is poor agreement between several structure-derived features calculated using different software packages or different versions and protocols of the same software, highlighting the need for transparency in algorithmic implementations and sharing of the structures used to aid reproducibility and comparison.[14] While some of the features (for example, the patch related features) calculated by PROPERMAB can be calculated using other software packages, the lack of access to the implementation details of these features in other software packages hinders reproducibility and comparison. This is further amplified by the broad spectrum of developability properties usually considered in antibody development and the potentially infinite number of molecular features that can be calculated from sequences and structures. We believe that the open-source nature of PROPERMAB will aid the comparison and development of new molecular features. One future direction for developing the package would be to represent antibody structures as molecular graphs to leverage the latest geometric deep learning (GDL) architectures for molecular property prediction.[63–65] While PROPERMAB's current Sequence Featurizer component calculates features based on full sequences, the Structure Featurizer module only calculates features derived from the Fv domain, as current structure prediction methods for immune proteins such as ABodyBuilder2[36] only predicts the 3D structure of the Fv domain. Whole-mAb structures and features derived from them are expected to further improve *in silico* prediction of developability properties.

While having the capability to calculate informative molecular features at large scale with low cost is critical, a major bottleneck in developing accurate ML models for developability prediction has been the lack of large and high-quality training data. Depending on the developability property being studied, the total number of datapoints in a training set can be as low as a few dozen (for example, viscosity).[25,26] The issue of lack of training data is further complicated by discrepancies in experimental data such as unmatched experimental conditions, different measuring instrumentations, and batch effects. Training models using aggregated datasets or applying models across datasets can easily violate the implicit assumption underlying many ML algorithms that the training data are independently and identically distributed. Such violations may partially explain the poor transferability of some recently developed developability prediction models.[14,25] Thus, targeted generation of large-scale experimental measurements under consistent and well-controlled conditions is desirable and expected to alleviate these issues.

Methods

Calculating sequence-based features

The theoretical_pi feature is calculated based on the Henderson-Hasselbalch equation by accounting for six amino acids D, E, H, K, R, and Y with ionizable side chains, and mAb terminal groups.[66] According to the Henderson-Hasselbalch equation, the net charge q of a mAb at a given pH is

$$q = \sum_{i}^{N} \frac{-1}{1 + 10^{\mathrm{pK_i - pH}}} + \sum_{j}^{P} \frac{1}{1 + 10^{\mathrm{pH - pK_j}}}$$

where i runs through all negatively charged amino acids (including tyrosine), j runs through all positively charged amino acids, N and P are the number of negatively and positively charged amino acids, respectively. To determine the theoretical pI, we consider the full mAb sequence, i.e., a pair of heavy chains and a pair of light chains, and the theoretical pI is determined as the pH that minimizes the square of the net charge q. For amino acid pK values, we used the pK set available at EMBOSS.[67]

We calculate n_charged_res by counting the total number of charged residues in full mAb sequence. At the default pH = 7.4, we consider residues D, E, K, R as charged. Similarly, n_charged_res_fv is computed by counting the total number of charged residues in the Fv domain. fv_charge is computed as the total charge of the Fv domain. At the default pH = 7.4, we assign residues D and E a −1 charge, and K and R a + 1 charge. fv_csp stands for the charge separation between the VH domain and VL domain and is defined as the total charge of the VH domain times the total charge of the VL domain. Similarly, fab_fc_csp stands for the charge separation between the Fab domain and the Fc domain and is defined as the total charge of the Fab domain times the total charge of the Fc domain. We define the Fab-Fc boundaries based on UniProt sequence annotations.[68] cdr_h3_length stands for the length of the CDR-H3 loop and is calculated by counting the number of residues in the CDR-H3 loop as defined according to the IMGT numbering scheme.[56]

Calculating structure-based features

All structure-based features are calculated based on 3D structures of the Fv domain, which are predicted by using the antibody-specific ABodyBuilder2 model of the ImmuneBuilder suite (version 0.0.8).[36] PROPERMAB uses the IMGT numbering scheme[56] by default because it is the default numbering scheme used by ABodyBuilder2 to number residues in predicted structures. However, features can also be calculated according to the Chothia[69] and Kabat[70] numbering schemes if explicitly specified. In general, we assigned atomic partial charges according to the CHARMM36 force field[71] using the molecular system and topology utilities from the OpenMM Python API (version 8.0).[72] Solvent accessibility of atoms and residues is computed using the FreeSASA package (version 2.1.0).[73] An atom is considered solvent exposed if its solvent accessibility is greater than 0, and a residue is considered solvent exposed if its relative solvent accessibility (RSA) is greater than or equal to 0.05. To calculate patch-related features, we first used NanoShaper (version 0.7.8)[74] to calculate a triangulated mesh representation of the molecular surface of the Fv structure with grid_scale = 0.5. We then used the APBS tool (version 3.0.0)[75] to calculate the Poisson-Boltzmann electrostatic potentials and the multivalue utility available as part of APBS tool to assign electrostatic potential at each vertex of the meshed surface. To assign hydrophobic potentials to mesh vertices, we implemented the algorithm by Heiden et al.[76,77] In the following, we describe the algorithms we implemented to calculate individual structure-based features.

net_charge and exposed_net_charge

The net_charge feature of the Fv domain is calculated by summing up the partial charges of all atoms in the domain. Similarly, the exposed_net_charge feature is calculated by summing up the partial charges of all solvent-exposed atoms.

net_charge_cdr and exposed_net_charge_cdr

The net_charge_cdr feature is the total charge in CDR regions. It is calculated by summing up the partial charges of all atoms located in CDR regions. The exposed_net_charge_cdr feature is the total charge of CDR atoms that are solvent-exposed and is calculated similarly.

scm

The scm feature, standing for the spatial charge map score, characterizes the overall magnitude of exposed negative electrostatics on the Fv domain.[29] In PROPERMAB, we calculate the scm feature according to the algorithm described in Agrawal et al.[29] Briefly, the scm scores for individual atoms are calculated as

$$scm_i = \sum_{j \in \mathcal{E}_i} q_j$$

where $\mathcal{E}_i$ is the set of side-chain atoms belonging to solvent-exposed residues and are within 10 Å of atom i. A residue is considered solvent-exposed if its side-chain solvent accessible surface area is >10 Å^2. The scm score for the Fv domain is then calculated as

$$scm = \left| \sum_{i \in \mathcal{A}} scm_i \times H(-scm_i) \right|$$

where $\mathcal{A}$ is the set of all atoms of the Fv domain and H is the Heaviside step function.

Fv_chml and exposed_FV_chml

The Fv_chml is a feature that describes the charge asymmetry between the VH domain and the VL domain. It is calculated by subtracting the net charge of the VL domain from the net charge of the VH domain. The exposed_Fv_chml feature describes the surface charge asymmetry between the VH domain and VL domain. It is calculated by subtracting the exposed net charge of the VL domain from that of the VH domain.

dipole_moment

Calculating the dipole moment of a protein is a complex topic.[78] However, the dipole moment for a collection of n point charges q_i at locations $\boldsymbol{r}_i$ can be calculated via a summation

$$\mathbf{p} = \sum_i^n q_i \boldsymbol{r}_i$$

In PROPERMAB we treat the charge distribution of the Fv domain as a collection of n point charges at atom locations and use the summation above to compute the dipole moment of the Fv domain. The locations $\boldsymbol{r}_i$ are centered at the center of geometry of the domain, that is $\boldsymbol{r}_i = \boldsymbol{r}_{i,0} - \boldsymbol{r}_0$, where $\boldsymbol{r}_{i,0}$ is the original coordinates in the PDB file, and

$$\boldsymbol{r}_0 = \frac{1}{n} \sum_i^n \boldsymbol{r}_{i,0}$$

is the center of geometry of the domain. We use the magnitude of the dipole moment as a feature. We note that the magnitude of dipole moment is multiplied by 4.803 to convert from Angstrom-electron-charge units to Debyes as described in Felder et al.[79]

aromatic_cdr and exposed_aromatic

This aromatic_cdr feature is calculated by counting the total number of aromatic residues (i.e., F, W, Y) across all CDR regions, which are defined according to the IMGT numbering scheme as previously described. The exposed_aromatic feature is calculated by counting the total number of aromatic residues in the Fv domain that are solvent exposed. A residue is solvent exposed if its RSA is 0.05 or bigger as previously described.

hyd_moment

The hyd_moment feature stands for hydrophobic moment, which is the analogue of the electric dipole moment for hydrophobicity. Just as the electric dipole moment measures the asymmetry of the charge distribution, the hydrophobic moment measures the amphiphilicity (asymmetry of hydrophobicity) of the structure. In PROPERMAB the hydrophobic moment is computed as

$$h = \sum_i^N h_i r_i$$

where h_i is the hydrophobicity of residue i, r_i is the center of geometry of residue i, and N is the total number of residues. PROPERMAB offers multiple options for amino acid hydrophobicity scales, the Kyle-Doolittle (KD) scale[80] is used by default.

heiden_score

The heiden_score feature is calculated as a weighted sum of the hydrophobic potentials of all molecular surface mesh vertices whose hydrophobic potential is positive as described in Waibl et al.[77] Briefly, the hydrophobic potential of vertex k, ϕ_k, is calculated as a weighted sum of the atomic lipophilicity of nearby atoms

$$\phi_k = \frac{\sum_{j \in \mathcal{N}(j,k)} g(r_{j,k}) l_j}{\sum_{j \in \mathcal{N}(j,k)} g(r_{j,k})}$$

where $\mathcal{N}(j,k)$ is the set of neighboring atoms of vertex k, $r_{j,k}$ is the Euclidean distance between atom j and vertex k, l_j is the lipophilicity of atom j as described in Wildman et al.,[81] and $g(r)$ is a function that assigns a distance-based weight and is defined as

$$g(r) = \frac{\exp\left(-\alpha \frac{R}{2}\right) + 1}{\exp\left(\alpha\left(r - \frac{R}{2}\right)\right) + 1}$$

where $\alpha = 1.5$ Å^{-1} and $R = 5$ Å, set according to Heiden et al.[76] To calculate the hidden_score feature, the hydrophobic potential of each vertex is weighted by its surface area s_k:

$$heiden\_score = \sum_k \max(\phi_k, 0)s_k$$

s_k is calculated by splitting the area of each triangle evenly to its three vertices.

hyd_asa, hph_asa, and aromatic_asa

The hyd_asa feature stands for the hydrophobic solvent accessible surface area. It is calculated by summing up the apolar component of the solvent accessible surface area of all residues of the Fv domain. Similarly, the hydrophilic solvent accessible surface area (hph_asa) is calculated by summing up the polar component of the solvent accessible surface area of all residues. Calculation of solvent accessible surface areas and their decomposition into polar and apolar components are done using FreeSASA, as previously described. The total solvent accessible aromatic surface area (aromatic_asa) is calculated by summing up the solvent accessible surface areas of aromatic residues.

pos_ann_index, neg_ann_index, and aromatic_ann_index

For a given set of residues, the Average Nearest Neighbor (ANN) metric measures the average distance between the residues and their nearest neighbors. If the average distance is less than the average for a hypothetical random distribution, then the distribution of the residues being analyzed is considered clustered. Otherwise, the residues are considered dispersed. In PROPERMAB we calculate the ratio of the observed average distance to the expected average distance and name this feature ann_index, i.e.

$$ann\_index = \frac{\overline{d_o}}{\overline{d_e}}$$

where $\overline{d_o}$ is the average of observed distance between each residue and its nearest neighbor

$$\overline{d_o} = \frac{\sum d_i}{N}$$

and $\overline{d_e}$ is the expected average distance for the set of residues under a null distribution and N is the size of the set. We simulate the null distribution by permuting the residues over locations of all solvent-exposed residues. Currently, PROPERMAB implements this feature for positively charged (pos_ann_index), negative charged (neg_ann_index), and aromatic residues (aromatic_ann_index), respectively.

pos_ripley_k, neg_ripley_k, and aromatic_ripley_k

The Ripley's K function is a spatial statistic that summarizes spatial dependence (residue clustering or dispersion) over a range of distances. For a given set of residues, the Ripley's K calculates the average number of neighboring residues associated with each residue at a given distance cutoff. If the average number of neighbors at the distance cutoff is higher than the average number of neighbors expected under a null distribution, then the residues are considered clustered. In PROPERMAB, the Ripley's K statistic (named ripley_k) is defined as

$$ripley\_k = \frac{k_o}{k_e}$$

where k_o is the observed proportion of neighboring residues evaluated at the given distance cutoff d_{cutoff}, i.e.

$$k_o = \frac{\sum_{i=1}^{N} \sum_{j,j \neq i}^{N} I\left(d_{i,j} < d_{cutoff}\right)}{N(N-1)}$$

where N is the total number of residues, $d_{i,j}$ is the Euclidean distance between residues i and j and I is the indicator function that evaluates to 1 when $d_{i,j}$ is less than d_{cutoff} and 0 otherwise. k_e is the expected proportion of neighboring residues for the evaluated distance under a null distribution. We simulate the null distribution by permuting the residues over the locations of all solvent-exposed residues. Currently, PROPERMAB implements this feature for positively charged (pos_ripley_k), negative charged (neg_ripley_k), and aromatic residues (aromatic_ripley_k), respectively.

hyd_patch_area, hyd_patch_area_cdr, pos_patch_area, pos_patch_area_cdr, neg_patch_area, neg_patch_area_cdr

To detect surface patches of a given biophysical property (positively charged, negatively charged, or hydrophobic), we first assign a property value to each triangle on the surface mesh by taking the average value of its three vertices. We then remove all triangles whose property value does not meet a given property threshold. For the remaining triangles, we employ the DBSCAN algorithm[82] as implemented in scikit-learn to cluster them based on the Cartesian coordinates of their geometric centers. Clusters of triangles whose total areas exceed a given area threshold are considered as surface patches. We set the area thresholds on positively charged and negatively charged patches at 20 Å^2, and the threshold on hydrophobic patches at 40 Å^2, determined through empirical evaluation (Supplementary Table S4) and visual inspection. A patch is considered near CDR if there is at least one vertex of the patch that is within 5 Å of any CDR vertices. NanoShaper assigns vertices to residues, and we define CDR vertices as the set of vertices that get assigned to the residues of CDR regions.

Sequence embeddings

PLMs have seen wide applications in protein property prediction and design.[83–85] PLMs transform textual protein sequences into contextual high-dimensional numerical representations called sequence embeddings that can be fed as input to downstream machine learning tasks. We implemented the SeqEmbedder class in PROPERMAB to facilitate the embedding of sequences using the ESM-1b[34] and AntiBERTy[37] models. We provide the options for extracting embeddings from PLMs as an additional feature for interested users, but did not use PLM embedding to generate any of the data discussed in this manuscript.

3D voxel grids

The 3D structures of proteins can be treated as if they were multi-channel 3D images known as 3D voxel grids.[86] Such representation makes protein structures amenable to 3D convolutional neural networks (3D-CNNs) and has enabled the application of 3D-CNNs in various protein-related tasks such as protein-ligand binding affinity prediction,[87] prediction of changes in thermodynamic stability,[88] protein design,[89] and more recently, prediction of antibody viscosity.[26] However, creating 3D voxel grid representations of protein structures can be a challenging task. In PROPERMAB, we implemented a simple interface that enables the transformation of protein 3D structure coordinates into 3D voxel grid representations with a single function call, while hiding the details of voxel creation and featurization. Such functionality and simple interface are expected to facilitate the use of 3D-CNNs for antibody developability prediction.

Training and evaluating ML models for predicting HIC RT and viscosity

We trained and evaluated the ML models as implemented in the scikit-learn package (version 1.2.2). For HIC RT prediction, our model-building procedure consists of wrapping feature standardization via a StandardScaler and parameter fitting of the ElasticNet model in a pipeline and using GridSearchCV to fit parameters and tune hyperparameters of the model in the pipeline. As an example, we searched a combination of the l1_ratio and the alpha parameters, with the l1_ratio ranging from 0.1 to 0.9 with a step size of 0.1, and four alpha values of 0.01, 0.05, 0.1, and 0.5. We then nested our GridSearchCV procedure within a leave-one-out cross-validation loop to estimate the generalization performance of our model-building procedure. Such a nested cross-validation is necessary to avoid overestimation of model generalization ability.[90] The RandomForestRegressor model for viscosity prediction is trained and evaluated by following the same procedure as HIC RT prediction. We selected the ElasticNet model primarily due to its interpretability while performance in predicting HIC retention time also turned out to be good. Results from another study suggest that prediction accuracy may be further improved by combining charge and aromaticity-related features into a non-linear model,[58] although at the expense of interpretability. For viscosity prediction, with the expectation that the relationships between the molecular features and viscosity are likely highly complex, we decided to use the RandomForestRegressor due to its ability to fit highly non-linear functions. This choice is further corroborated by the fact that none of the features showed strong linear correlation with viscosity.

Acknowledgments

The authors would like to thank Robert Babb and Calvin Chen from Regeneron Protein Expression Sciences for their discussions and suggestions during the project. The authors also thank Eric Prager from Regeneron Research Program Management for his administrative support.

Disclosure statement

B.L., S.K.L., W.H.W., J.H.X., D.J.L., M.S., J.M., and M.C.F. are employees of Regeneron Pharmaceuticals and may hold company stock and/or stock options. M.S. is an officer of Regeneron Pharmaceuticals. P.G.H. was an employee of Regeneron Pharmaceuticals at the time of the studies and may hold company stock and/or stock options. P.G.H. has subsequently become a Databricks employee and equity holder at time of submission. G.S.A. is an employee of Flagship Pioneering and is a scientific advisory board member of Micoy Therapeutics.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Matthew C. Franklin http://orcid.org/0000-0002-1482-0136

Code availability

The source code of PROPERMAB is available at https://github.com/regeneron-mpds/propermab. for noncommercial uses.

Data availability statement

The molecular features computed using PROPERMAB for the HIC RT dataset and the Ab21 and PDGF38 viscosity datasets are available as Supplementary Datasets.

References

1. Lu RM, Hwang Y-C, Liu I-J, Lee C-C, Tsai H-Z, Li H-J, Wu H-C. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020;27(1):1. doi:10.1186/s12929-019-0592-z.
2. Nelson AL, Dhimolea E, Reichert JM. Development trends for human monoclonal antibody therapeutics. Nat Rev Drug Discov. 2010;9(10):767–774. doi:10.1038/nrd3229.
3. Crescioli S, Kaplon H, Chenoweth A, Wang L, Visweswaraiah J, Reichert JM. Antibodies to watch in 2024. MAbs. 2024;16(1):2297450. doi:10.1080/19420862.2023.2297450.
4. Scott AM, Wolchok JD, Old LJ. Antibody therapy of cancer. Nat Rev Cancer. 2012;12(4):278–287. doi:10.1038/nrc3236.
5. Zinn S, Vazquez-Lombardi R, Zimmermann C, Sapra P, Jermutus L, Christ D. Advances in antibody-based therapy in oncology. Nat Cancer. 2023;4(2):165–180. doi:10.1038/s43018-023-00516-z.
6. Focosi D, McConnell S, Casadevall A, Cappello E, Valdiserra G, Tuccori M. Monoclonal antibody therapies against SARS-CoV-2. Lancet Infect Dis. 2022;22(11):e311–e326. doi:10.1016/S1473-3099(22)00311-5.
7. Mieczkowski C, Zhang X, Lee D, Nguyen K, Lv W, Wang Y, Zhang Y, Way J, Gries J-M. Blueprint for antibody biologics developability. MAbs. 2023;15(1):2185924. doi:10.1080/19420862.2023.2185924.
8. Bauer J, Rajagopal N, Gupta P, Gupta P, Nixon AE, Kumar S. How can we discover developable antibody-based biotherapeutics? Front Mol Biosci. 2023;10:1221626. doi:10.3389/fmolb.2023.1221626.
9. Xu Y, Wang D, Mason B, Rossomando T, Li N, Liu D, Cheung JK, Xu W, Raghava S, Katiyar A, et al. Structure, heterogeneity and developability assessment of therapeutic antibodies. MAbs. 2019;11(2):239–264. doi:10.1080/19420862.2018.1553476.

10. Hong T, Iwashita K, Shiraki K. Viscosity control of protein solution by Small solutes: a review. Curr Protein Pept Sci. 2018;19(8):746–758. doi:10.2174/1389203719666171213114919.
11. Zarzar J, Khan T, Bhagawati M, Weiche B, Sydow-Andersen J, Sreedhara A. High concentration formulation developability approaches and considerations. MAbs. 2023;15(1):2211185. doi:10.1080/19420862.2023.2211185.
12. Evers A, Krah S, Demir D, Gaa R, Elter D, Schroeter C, Zielonka S, Rasche N, Dotterweich J, Knuehl C, et al. Engineering hydrophobicity and manufacturability for optimized biparatopic antibody–drug conjugates targeting c-*MET*. MAbs. 2024;16(1):2302386. doi:10.1080/19420862.2024.2302386.
13. Fernandez-Quintero ML, Ljungars A, Waibl F, Greiff V, Andersen JT, Gjølberg TT, Jenkins TP, Voldborg BG, Grav LM, Kumar S, et al. Assessing developability early in the discovery process for novel biologics. MAbs. 2023;15(1):2171248. doi:10.1080/19420862.2023.2171248.
14. Jain T, Boland T, Vasquez M. Identifying developability risks for clinical progression of antibodies using high-throughput in vitro and in silico approaches. MAbs. 2023;15(1):2200540. doi:10.1080/19420862.2023.2200540.
15. Khetan R, Curtis R, Deane CM, Hadsund JT, Kar U, Krawczyk K, Kuroda D, Robinson SA, Sormanni P, Tsumoto K, et al. Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. MAbs. 2022;14(1):2020082. doi:10.1080/19420862.2021.2020082.
16. Feng J, Jiang M, Shih J, Chai Q. Antibody apparent solubility prediction from sequence by transfer learning. iScience. 2022;25(10):105173. doi:10.1016/j.isci.2022.105173.
17. Wolf Perez AM, Sormanni P, Andersen JS, Sakhnini LI, Rodriguez-Leon I, Bjelke JR, Gajhede AJ, De Maria L, Otzen DE, Vendruscolo M, et al. In vitro and in silico assessment of the developability of a designed monoclonal antibody library. MAbs. 2019;11(2):388–400. doi:10.1080/19420862.2018.1556082.
18. Makowski EK, Wang T, Zupancic JM, Huang J, Wu L, Schardt JS, De Groot AS, Elkins SL, Martin WD, Tessier PM. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. Nat Biomed Eng. 2023;8(1):45–56. doi:10.1038/s41551-023-01074-6.
19. Lai PK, Gallegos A, Mody N, Sathish HA, Trout BL. Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics. MAbs. 2022;14(1):2026208. doi:10.1080/19420862.2022.2026208.
20. Park E, Izadi S. Molecular surface descriptors to predict antibody developability: sensitivity to parameters, structure models, and conformational sampling. MAbs. 2024;16(1):2362788. doi:10.1080/19420862.2024.2362788.
21. Harmalkar A, Rao R, Richard Xie Y, Honer J, Deisting W, Anlahr J, Hoenig A, Czwikla J, Sienz-Widmann E, Rau D, et al. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. MAbs. 2023;15(1):2163584. doi:10.1080/19420862.2022.2163584.
22. Waight AB, Prihoda D, Shrestha R, Metcalf K, Bailly M, Ancona M, Widatalla T, Rollins Z, Cheng AC, Bitton DA, et al. A machine learning strategy for the identification of key in silico descriptors and prediction models for IgG monoclonal antibody developability properties. MAbs. 2023;15(1):2248671. doi:10.1080/19420862.2023.2248671.
23. Mock M, Jacobitz AW, Langmead CJ, Sudom A, Yoo D, Humphreys SC, Alday M, Alekseychyk L, Angell N, Bi V, et al. Development of in silico models to predict viscosity and mouse clearance using a comprehensive analytical data set collected on 83 scaffold-consistent monoclonal antibodies. MAbs. 2023;15(1):2256745. doi:10.1080/19420862.2023.2256745.
24. Hoerschinger VJ, Waibl F, Pomarici ND, Loeffler JR, Deane CM, Georges G, Kettenberger H, Fernández-Quintero ML, Liedl KR. PEP-Patch: electrostatics in protein–protein recognition, specificity, and Antibody Developability. J Chem Inf Model. 2023;63(22):6964–6971. doi:10.1021/acs.jcim.3c01490.
25. Lai PK, Fernando A, Cloutier TK, Gokarn Y, Zhang J, Schwenger W, Chari R, Calero-Rubio C, Trout BL. Machine learning applied to determine the molecular descriptors responsible for the viscosity behavior of concentrated therapeutic antibodies. Mol Pharm. 2021;18(3):1167–1175. doi:10.1021/acs.molpharmaceut.0c01073.
26. Rai BK, Apgar JR, Bennett EM. Low-data interpretable deep learning prediction of antibody viscosity using a biophysically meaningful representation. Sci Rep. 2023;13(1). doi:10.1038/s41598-023-28841-4.
27. Makowski EK, Chen H-T, Wang T, Wu L, Huang J, Mock M, Underhill P, Pelegri-O'Day E, Maglalang E, Winters D, et al. Reduction of monoclonal antibody viscosity using interpretable machine learning. MAbs. 2024;16(1):2303781. doi:10.1080/19420862.2024.2303781.
28. Thorsteinson N, Gunn JR, 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. doi:10.1080/19420862.2021.1981805.
29. Agrawal NJ, Helk B, Kumar S, Mody N, Sathish HA, Samra HS, Buck PM, Li L, Trout BL. Computational tool for the early screening of monoclonal antibodies for their viscosities. MAbs. 2016;8(1):43–48. doi:10.1080/19420862.2015.1099773.
30. Sankar K, Krystek SR, Carl SM, Day T, Maier JKX. AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins. 2018;86(11):1147–1156. doi:10.1002/prot.25594.
31. Sharma VK, Patapoff TW, Kabakoff B, Pai S, Hilario E, Zhang B, Li C, Borisov O, Kelley RF, Chorny I, et al. In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. Proc Natl Acad Sci UA. 2014;111(52):18601–18606. doi:10.1073/pnas.1421779112.
32. Li L, Kumar S, Buck PM, Burns C, Lavoie J, Singh SK, Warne NW, Nichols P, Luksha N, Boardman D. Concentration dependent viscosity of monoclonal antibody solutions: explaining experimental behavior in terms of molecular properties. Pharm Res. 2014;31(11):3161–3178. doi:10.1007/s11095-014-1409-0.
33. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589. doi:10.1038/s41586-021-03819-2.
34. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15). doi:10.1073/pnas.2016239118.
35. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–876. doi:10.1126/science.abj8754.
36. Abanades B, Wong WK, Boyles F, Georges G, Bujotzek A, Deane CM. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. Commun Biol. 2023;6(1):575. doi:10.1038/s42003-023-04927-7.
37. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. Machine learning for structural biology workshop, NeurIPS. 2021. https://arxiv.org/abs/2112.07782.
38. Pedregosa F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
39. Grinshpun B, Thorsteinson N, Pereira JN, Rippmann F, Nannemann D, Sood VD, Fomekong Nanfack Y. Identifying biophysical assays and in silico properties that enrich for slow

clearance in clinical-stage therapeutic antibodies. MAbs. 2021;13 (1):1932230. doi:10.1080/19420862.2021.1932230.

40. Ausserwöger H, Krainer G, Welsh TJ, Thorsteinson N, de Csilléry E, Sneideris T, Schneider MM, Egebjerg T, Invernizzi G, Herling TW, et al. Surface patches induce nonspecific binding and phase separation of antibodies. Proc Natl Acad Sci USA. 2023;120 (15):e2210332120. doi:10.1073/pnas.2210332120.
41. Ausserwöger H, Schneider MM, Herling TW, Arosio P, Invernizzi G, Knowles TPJ, Lorenzen N. Non-specificity as the sticky problem in therapeutic antibody development. Nat Rev Chem. 2022;6(12):844–861. doi:10.1038/s41570-022-00438-x.
42. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. Boca Raton: Chapman and Hall/CRC Press; 2015.
43. Sivley RM, Dou X, Meiler J, Bush WS, Capra JA. Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. Am J Hum Genet. 2018;102 (3):415–426. doi:10.1016/j.ajhg.2018.01.017.
44. Rego NB, Xi E, Patel AJ. Identifying hydrophobic protein patches to inform protein interaction interfaces. Proc Natl Acad Sci U S A. 2021;118(6). doi:10.1073/pnas.2018234118.
45. McCue JT. Theory and use of hydrophobic interaction chromatography in protein purification applications. Methods Enzymol. 2009;463:405–414.
46. Haverick M, Mengisen S, Shameem M, Ambrogelly A. Separation of mAbs molecular variants by analytical hydrophobic interaction chromatography HPLC: overview and applications. MAbs. 2014;6 (4):852–858. doi:10.4161/mabs.28693.
47. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci UA. 2017;114(5):944–949. doi:10.1073/pnas.1616408114.
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B: Stat Methodol. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
49. Wang X, Das TK, Singh SK, Kumar S. Potential aggregation prone regions in biotherapeutics: a survey of commercial monoclonal antibodies. MAbs. 2009;1(3):254–267. doi:10.4161/mabs.1.3.8035.
50. Shire SJ, Shahrokh Z, Liu J. Challenges in the development of high protein concentration formulations. J Pharm Sci. 2004;93 (6):1390–1402. doi:10.1002/jps.20079.
51. Apgar JR, Tam ASP, Sorm R, Moesta S, King AC, Yang H, Kelleher K, Murphy D, D'Antona AM, Yan G, et al. Modeling and mitigation of high-concentration antibody viscosity through structure-based computer-aided protein design. PLOS ONE. 2020;15(5):e0232713. doi:10.1371/journal.pone.0232713.
52. Lai PK, Ghag G, Yu Y, Juan V, Fayadat-Dilman L, Trout BL. Differences in human IgG1 and IgG4 S228P monoclonal antibodies viscosity and self-interactions: experimental assessment and computational predictions of domain interactions. MAbs. 2021;13 (1):1991256. doi:10.1080/19420862.2021.1991256.
53. Tomar DS, Kumar S, Singh SK, Goswami S, Li L. Molecular basis of high viscosity in concentrated antibody solutions: strategies for high concentration drug product development. MAbs. 2016;8 (2):216–228. doi:10.1080/19420862.2015.1128606.
54. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (NY). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.
55. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. 2021;31(1):141–146. doi:10.1002/pro.4205.
56. Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and ig superfamily V-like domains. Dev Comp Immunol. 2003;27(1):55–77. doi:10.1016/S0145-305X(02)00039-3.
57. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. doi:10.1093/bioinformatics/btv552.
58. Hebditch M, Warwicker J. Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. PeerJ. 2019;7:e8199. doi:10.7717/peerj.8199.
59. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci UA. 2014;111 (9):3354–3359. doi:10.1073/pnas.1309933111.
60. Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci UA. 2019;116(10):4025–4030. doi:10.1073/pnas.1810576116.
61. Lai PK. DeepSCM: an efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. Comput Struct Biotechnol J. 2022;20:2143–2152. doi:10.1016/j.csbj.2022.04.035.
62. Kalejaye L, Wu I-E, Terry T, Lai P-K. DeepSP: deep learning-based spatial properties to predict monoclonal antibody stability. Comput Struct Biotechnol J. 2024;23:2220–2229. doi:10.1016/j.csbj.2024.05.029.
63. Bronstein MM, Bruna, J., Cohen, T., Veličković, P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. arXiv Prepr arXiv: 2104.13478. 2021. https://arxiv.org/abs/2104.13478.
64. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nat Mach Intel. 2021;3(12):1023–1032. doi:10.1038/s42256-021-00418-8.
65. Velickovic P. Everything is connected: graph neural networks. Curr Opin Struct Biol. 2023;79:102538. doi:10.1016/j.sbi.2023.102538.
66. Kozlowski LP. IPC – isoelectric point calculator. Biol Direct. 2016;11(1):55. doi:10.1186/s13062-016-0159-9.
67. EMBOSS. Amino acid pK data. [cited 2024 May]; Available from: https://emboss.sourceforge.net/apps/cvs/emboss/apps/iep.html.
68. UniProt C, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bye-A-Jee H, Cukura A, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–D531. doi:10.1093/nar/gkac1052.
69. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. J Mol Biol. 1997;273 (4):927–948. doi:10.1006/jmbi.1997.1354.
70. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med. 1970;132(2):211–250. doi:10.1084/jem.132.2.211.
71. Huang J, MacKerell AD Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. J Comput Chem. 2013;34(25):2135–2145. doi:10.1002/jcc.23354.
72. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang L-P, Simmonett AC, Harrigan MP, Stern CD, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLOS Comput Biol. 2017;13(7):e1005659. doi:10.1371/journal.pcbi.1005659.
73. Mitternacht S. FreeSASA: an open source C library for solvent accessible surface area calculations. F1000Res. 2016;5:189. doi:10.12688/f1000research.7931.1.
74. Decherchi S, Rocchia W, Lorenz C. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. PLOS ONE. 2013;8(4):e59744. doi:10.1371/journal.pone.0059744.
75. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, et al. Improvements to

the APBS biomolecular solvation software suite. Protein Sci. 2018;27(1):112–128. doi:10.1002/pro.3280.

76. Heiden W, Moeckel G, Brickmann J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. J Comput Aided Mol Des. 1993;7(5):503–514. doi:10.1007/BF00124359.
77. Waibl F, Fernández-Quintero ML, Wedl FS, Kettenberger H, Georges G, Liedl KR. Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. Front Mol Biosci. 2022;9:960194. doi:10.3389/fmolb.2022.960194.
78. Antosiewicz J. Computation of the dipole moments of proteins. Biophys J. 1995;69(4):1344–1354. doi:10.1016/S0006-3495(95)80001-9.
79. Felder CE, Prilusky J, Silman I, Sussman JL. A server and database for dipole moments of proteins. Nucleic Acids Res. 2007;35(Web Server issue):W512–21. doi:10.1093/nar/gkm307.
80. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–132. doi:10.1016/0022-2836(82)90515-0.
81. Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comput Sci. 1999;39(5):868–873. doi:10.1021/ci990307l.
82. Ester M, Kriegel, H. P., Sander, J., Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96(34):226–231.
83. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J. 2021;19:1750–1758. doi:10.1016/j.csbj.2021.03.022.
84. Ferruz N, Höcker B. Controllable protein design with language models. Nat Mach Intel. 2022;4(6):521–532. doi:10.1038/s42256-022-00499-z.
85. Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst. 2021;12(6):654–669 e3. doi:10.1016/j.cels.2021.05.017.
86. Torng W, Altman RB. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinf. 2017;18(1):302. doi:10.1186/s12859-017-1702-0.
87. Jimenez J, Škalič M, Martínez-Rosell G, De Fabritiis G. K DEEP: protein–ligand absolute binding affinity prediction via 3D-Convolutional neural networks. J Chem Inf Model. 2018;58(2):287–296. doi:10.1021/acs.jcim.7b00650.
88. Li B, Yang YT, Capra JA, Gerstein MB. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLOS Comput Biol. 2020;16(11):e1008291. doi:10.1371/journal.pcbi.1008291.
89. Zhang Y, Chen Y, Wang C, Lo C-C, Liu X, Wu W, Zhang J. ProDCoNN: protein design using a convolutional neural network. Proteins. 2020;88(7):819–829. doi:10.1002/prot.25868.
90. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLOS ONE. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365.