molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

Machine Learning Models for Predicting Monoclonal Antibody Biophysical Properties from Molecular Dynamics Simulations and Deep Learning-Based Surface Descriptors

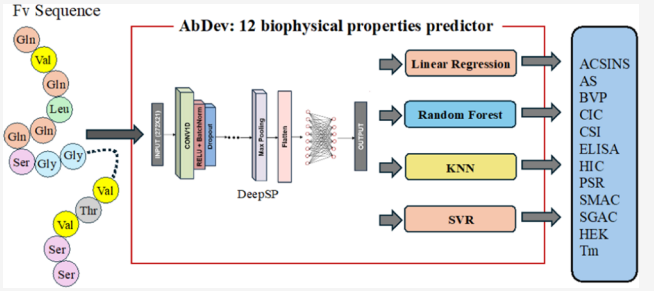
I-En Wu, Lateefat Kalejaye, and Pin-Kuang Lai*

Cite This: *Mol. Pharmaceutics* 2025, 22, 142−153

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Monoclonal antibodies (mAbs) have found extensive applications and development in treating various diseases. From the pharmaceutical industry's perspective, the journey from the design and development of mAbs to clinical testing and large-scale production is a highly time-consuming and resource-intensive process. During the research and development phase, assessing and optimizing the developability of mAbs is of paramount importance to ensure their success as candidates for therapeutic drugs. The critical factors influencing mAb development are their biophysical properties, such as aggregation propensity, solubility, and viscosity. This study utilized a data set comprising 12 biophysical properties of 137 antibodies from a previous study (Proc Natl Acad Sci USA. 114(5):944−949, 2017). We employed full-length antibody molecular dynamics simulations and machine learning techniques to predict experimental data for these 12 biophysical properties. Additionally, we utilized a newly developed deep learning model called DeepSP, which directly predicts the dynamical and structural properties of spatial aggregation propensity and spatial charge map in different antibody regions from sequences. Our research findings indicate that the machine learning models we developed outperform previous methods in predicting most biophysical properties. Furthermore, the DeepSP model yields similar predictive results compared to molecular dynamic simulations while significantly reducing computational time. The code and parameters are freely available at https://github.com/Lailabcode/AbDev. Also, the webapp, AbDev, for 12 biophysical properties prediction has been developed and provided at https://devpred.onrender.com/AbDev.



KEYWORDS: *machine learning, deep learning, molecular dynamics simulation, monoclonal antibody, developability*

INTRODUCTION

In recent years, monoclonal antibodies (mAbs) have become widely utilized to treat diverse diseases, including cancer, autoimmune disorders, and infectious diseases.[1,2] As of 2022, the global mAbs market size surpassed USD 200 billion, with expectations of USD 300 billion by 2025.[3] However, developing mAbs from initial design to FDA approval is time-consuming and costly. From the early stages of design through clinical trials to final production, discovering that a particular mAb candidate is unsuitable for therapeutic use can result in significant resource waste. Therefore, employing protein computational simulations in the mAb development process is desired to mitigate the time and cost of developing antibody drugs.[4−6]

Computational modeling and simulation, encompassing a diverse range of techniques, are pivotal in understanding, predicting, and optimizing various aspects of mAbs. These computational methods are crucial throughout the design, development, and analysis stages of mAbs, with molecular dynamics (MD) simulation standing out as a key technique in drug discovery and design processes.[7,8] MD simulation plays an essential role in identifying binding sites, refining virtual screening methods, and predicting ligand binding energies, thereby facilitating a deeper understanding of mAbs' behavior and interactions.[9,10] Brandt et al. demonstrated the application of MD simulation in tandem with continuum hydrodynamics modeling and experimental diffusion measurements to validate the conformational and hydrodynamic behavior of human Immunoglobulins G1 (IgG1) mAbs in aqueous solutions.[11] Similarly, Zamolo et al. explored the interaction of supported affinity ligands with mAbs through MD simulation, highlighting its utility in probing the nuances of ligand-mAb interactions.[12] Lapelosa et al. employed MD simulations to elucidate the mAb-mAb association, shedding light on the aggregation process of mAbs and contributing to the development of strategies to

Received: July 22, 2024
Revised: November 21, 2024
Accepted: November 22, 2024
Published: November 28, 2024

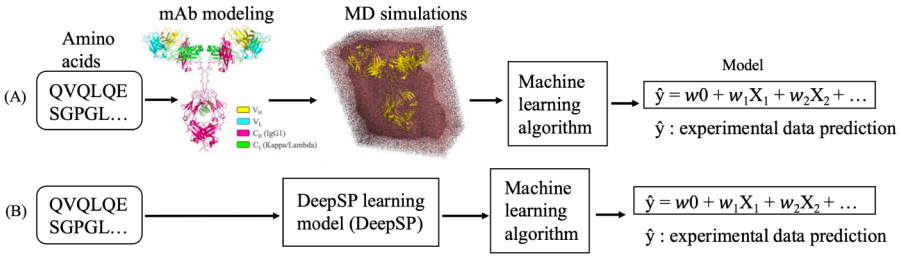


Figure 1. Flowchart of the research methods, using (A) MD simulations and (B) deep learning model (DeepSP) to generate the features for machine learning models training.

mitigate unwanted aggregation.[9] The integration of laboratory techniques with MD simulations presents a holistic approach to understanding the intricate protein−protein interactions within concentrated mAbs solutions, underlining the indispensable role of computational methods in advancing mAbs research and development.

Molecular modeling enables the prediction of mAbs structures when experimental structures are not available.[13] Several mAb modeling tools have been recently developed for three-dimensional structure prediction. ABodyBuilder,[14] an advanced and automated antibody modeling pipeline, enhances the prediction accuracy of antibody structures. It provides data-driven accuracy estimates and identifies potential sequence liabilities, facilitating further studies for mAbs. Phillips et al. offered comprehensive guidelines for mAbs using molecular modeling and MD simulations, providing a valuable protocol for understanding the protein modeling and the formulation of mAbs.[15] Computational modeling of mAbs can bridge the gap between experimental methods to probe the microscopic and transient molecular interactions. These computational approaches accelerate the process of mAb development by offering insights and predictions that complement experimental findings.

Machine learning has been increasingly applied in various stages of mAb development, from discovery to clinical development. Derek et al. proposed a method for predicting the antigen specificity of antibodies from their sequence using deep learning.[16] Lai et al. implemented machine learning to predict mAb aggregation and viscosity for high-concentration formulation development, using solvent-accessible surface area (SASA),[17] spatial aggregation propensity (SAP),[18] and spatial charge map (SCM)[19] derived from MD simulations as features.[20,21] In addition, the TAP: Therapeutic Antibody Profiler web server was utilized to predict the antibody developability.[22] Machine learning techniques can also be used to predict the pharmacokinetics and pharmacodynamics properties of mAbs, helping to understand how mAbs behave in the human body.[23,24] In short, machine learning approaches provide a breadth of applications in mAb development, highlighting its role in accelerating discovery, improving safety and efficacy, and reducing the time and cost of bringing new therapies to market.

Jain et al. prepared 137 clinical-stage antibodies, including 48 approved for therapeutic use, to construct isotype-matched IgG1 antibodies.[25] They assessed 12 biophysical property assays that provided insights into the distribution of biophysical metrics relevant to the "developability" of antibodies. The term "developability" for mAbs refers to the set of criteria and assessments used to evaluate and optimize the physical, chemical, and biological properties of mAbs to ensure they are stable for development as therapeutic agents. This includes their manufacturability, stability, solubility, safety, and efficacy. Jain's data set has been utilized by several studies for various applications. For instance, a sequence-based tool driven from standup monolayer adsorption chromatography (SMAC), salt-gradient affinity-capture self-interaction nanoparticle spectroscopy (SGAC-SINS), and hydrophobic interaction chromatography (HIC), called SSH[26] was developed to predict the hydrophobicity of mAb. Furthermore, structure-based predictive models were devised to assess two crucial biophysical properties: hydrophobicity and polyspecificity. Waibl et al. introduced a purely physics-based approach to predict the hydrophobic behavior of mAbs, relying on a localized description of the free energy of hydration.[27] In addition, Hebditch and Warwicker utilized 35 sequence features derived from the variable regions (Fv) of mAbs, including the standard 20 amino acid propensities, 7 composite scores of amino acids, and additional features like folding and disorder propensities, beta-strand propensities, and sequence entropy.[28] They developed machine learning models based on these sequence features to predict 12 biophysical properties of 137 mAbs.

In this study, we collected the Fv sequences for the 137 mAbs and experimental data for 12 biophysical properties obtained from Jain et al. We aim to develop predictive models for these 12 properties using machine learning approaches. We employed two methods to construct the predictive models, as depicted in Figure 1. In Figure 1A, full-length mAb structures were built using the homology model of the Fv structures from ABodyBuilder and a template full-length IgG1 structure. The features, SCM and SAP (The formulas are provided in the Supplementary document), required for training the machine learning models were obtained and calculated from MD simulations. However, this approach is not suitable for high-throughput analysis due to the time-consuming MD simulations of the full-length structures needed to obtain features for machine learning training. To address this issue, we utilized a deep learning model, DeepSP,[29] developed by our group to predict the SAP and SCM scores directly from the Fv sequences (Figure 1B). In this study, we built 12 predictive models for these biophysical properties from the MD-based and the DeepSP features, resulting in 24 models. We compared the performance of these models with published papers and found that our models exhibit improved performance. The codes, parameters, and models are freely available at https://github.com/Lailabcode/AbDev. In addition, the webapp, AbDev, for 12 biophysical properties prediction has been developed and deployed on https://devpred.onrender.com/AbDev.

MATERIALS AND METHODS

Data Collection. Jain et al.[25] provided a panel of 12 biophysical properties that are related to antibody develop-

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

ability, including (1) HEK Titer, (2) Fab melting temperature by differential scanning fluorimetry (DSF/Tm), (3) salt-gradient affinity-capture self-interaction nanoparticle spectroscopy (SGAC), (4) hydrophobic interaction chromatography (HIC), (5) standup monolayer adsorption chromatography (SMAC), (6) slope of accelerated stability (AS), (7) poly specificity reagent (PSR), (8) affinity-capture self-interaction nanoparticle spectroscopy (ACSINS), (9) cross-interaction chromatography (CIC), (10) clone self-interaction by biolayer interferometry (CSI), (11) ELISA and (12) baculovirus particle (BVP) have been recently published for 137 clinical to marketed mAbs. ACSINS and CSI are used to assess antibody self-interaction. PSR, BVP, CIC, and ELISA are measured to evaluate cross-interaction (off-target binding). Other parameters are commonly used biophysical characterization metrics in the lab. In the data set of 137 mAbs, light chain types were categorized into kappa and lambda subtypes. Specifically, 124 of the mAbs had kappa light chains, while 13 of the mAbs had lambda light chains. Due to this significant imbalance in the distribution of the kappa and lambda subtypes, it was challenging to come to a meaningful conclusion regarding potential correlations between light chain type and biophysical properties.

Data Preprocessing. From the data set provided by Jain et al., we observed that uneven distribution of certain experimental data can lead to the overfitting of machine learning models.[30] Training machine learning models on highly unevenly distributed data may result in issues predicting data that is either too large or too small in subsequent analyses. First, we removed some of the outliers from the data set. For instance, in the context of HIC, which measures the retention time of mAb in columns to calculate their hydrophobicity, two mAbs, lirilumab and nimotuzumab, were unable to flow out of the column within the designated time during the experiments. As a result, these two data points were removed from the data set. Therefore, we utilized the remaining 135 data points to establish predictive models for HIC. Similarly, for AS, CIC, and SMAC, we removed 1, 2, and 3 data points, respectively. To make our data more closely resemble a normal distribution, we utilized the *QuantileTransformer()* function from the sklearn.preprocessing module. This transformation was applied to the data from SMAC, CIC, ACSINS, and CSI, ensuring that their distributions followed a normal distribution. As for SGAC, we calculated its mean and standard deviation and then normalized the original data by subtracting the mean and dividing it by the standard deviation. While this approach may result in values that no longer hold direct physical meaning, the predicted values ultimately remain comparable. The results of the experimental data preprocessing are shown in Figure S1.

Molecular Modeling of mAb. In this research, full-length sequence modeling of 137 mAbs was constructed for the following MD simulations. The mAb molecules were built according to the methodology outlined by Brandt et al.[11] The Fv structures of the 137 mAbs were generated through homology modeling using ABodyBuilder, and these Fv regions were then aligned onto a full-length IgG1 template, derived from the KOL/Padlan[31,32] model that matches the light chain types to construct the complete mAb structures. The information on the Fv and the type of light chains were provided in Jain's work, and the glycosylation patterns specific to each mAb were then modeled employing the G0F structure. Also, we used the IMGT numbering definition to assign residue positions in mAbs, allowing for precise identification of disulfide bond locations. For example, in abituzumab, the disulfide bond positions are H22−H96, H'22-H'96, H145−H201, H'145-H'201, H221-L214, H227−H'227, H230−H'230, H262−H322, H'262-H'322, H368−H426, H'368-H'426, L23-L88, L'23-L'88, L135-L195, L'135-L'195 following IMGT numbering system. Here, H' and L' represent the disulfide bonds in the other heavy and light chains, respectively. These disulfide bonds were modeled to maintain structural integrity in both the heavy and light chains, as well as across the heavy chain domains. The IgG1 constant region served as the template for modeling, ensuring an accurate representation of the disulfide bonds during molecular dynamics simulations.

Molecular Dynamics (MD) Simulations. MD simulations were set up using all-atom structures with explicit solvents, employing the TIP3P water model. We used VMD to place a single antibody in a water box[33] extending 12 Å beyond the protein surface.[34] We implemented the NPT ensemble to maintain the temperature and pressure at 300 K and 1 atm, respectively, using the NAMD software package with the CHARMM36m force field.[35−37] To mimic the experimental conditions, the system pH value was adjusted to 7.3 using PROPKA3 protocol to assign the protonation states of histidine residues. Electrostatic interactions were calculated via the Particle Mesh Ewald (PME) method, while van der Waals interactions were computed with a switching distance of 10 Å and a cutoff of 12 Å. Integration time step was set to 2 fs. Preceding production runs, each mAb system underwent a 10 ns pre-equilibrium phase, followed by 50 ns of production runs. The production results were then used to calculate the SAP and SCM as the features for machine learning model training.

Deep Learning Model (DeepSP) for SAP and SCM Prediction. In this work, due to the extensive computational time and resources required for obtaining features necessary for training machine learning models using MD simulations, we have also employed a deep learning tool called DeepSP, developed from our previous work.[29] DeepSP can directly predict the SCM and SAP values for each domain of mAb solely by providing the variable region amino acid sequence. By following the instructions provided at https://github.com/Lailabcode/DeepSP, we can obtain the 30 SAP or SCM values in different regions in seconds for all 137 mAbs. In Figure S2, we compared the results calculated from MD simulations and DeepSP. Our findings indicate that DeepSP can predict the SAP and SCM values with both speed and accuracy.

Features Selection for Machine Learning. For the machine learning training, we utilized the SAP and SCM values extracted by DeepSP and MD simulation for each region, resulting in a total of 30 features. However, due to the limited data set size, there is a risk of overfitting when dealing with numerous features. Therefore, our initial step was to reduce the number of features used for model training.

In this study, our objective was to perform feature selection for predictive modeling using various regression algorithms in conjunction with the *ExhaustiveFeatureSelector()* from the mlxtend library.[38] We systematically evaluated different feature subsets based on the negative mean squared error as the scoring metric, varying the number of features and cross-validation folds. Subsequently, we calculated the mean squared error (MSE) for specific feature subsets identified by the Exhaustive Feature Selector (EFS). For each subset, MSE was computed using different regression models within a repeated 4-fold cross-validation framework.

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

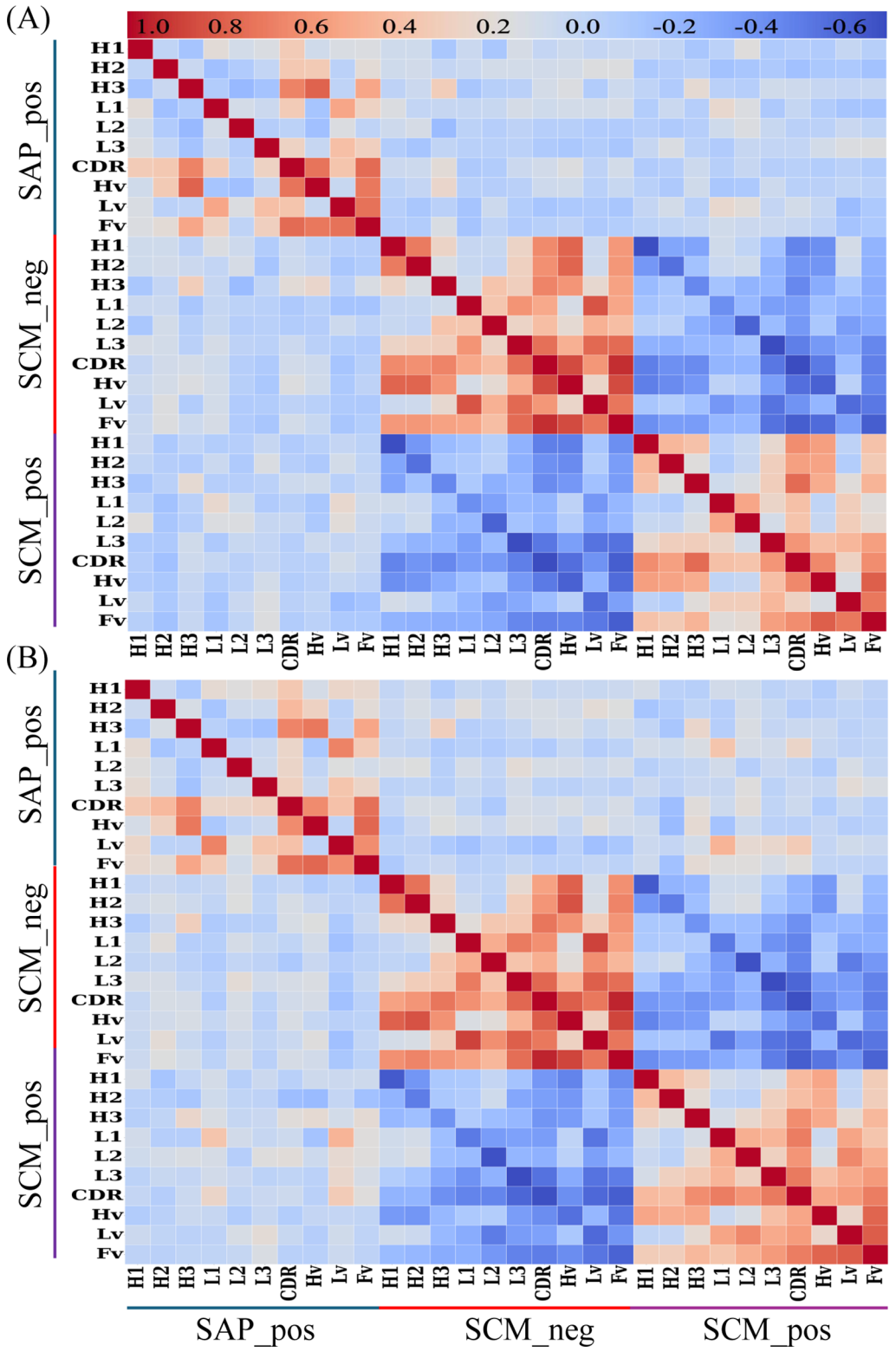


Figure 2. Correlation heatmap of features from (A) MD simulations and (B) DeepSP across antibody domains for machine learning models training. H1 refers to CDRH1, H2 to CDRH2, H3 to CDRH3, L1 to CDRL1, L2 to CDRL2, and L3 to CDRL3. The correlation values represented in the heatmaps are provided in the supplementary file: HeatMap_SI.xlsx.

Finally, we compiled detailed information on all subsets and their associated averaged MSE values, ultimately selecting the feature combination with the smallest MSE value to train the machine learning model. By integrating feature selection with machine learning model evaluation using EFS and repeated k-fold cross-validation, our approach enables the identification of optimal feature subsets and the assessment of model performance. This iterative process enhances the robustness and interpretability of regression models, facilitating informed decision-making in predictive modeling tasks.

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

Table 1. Best 3-Feature and 4-Feature Combinations From 4-Fold Cross-Validation with the Lowest MSE for ELISA Using DeepSP Features[a]

Regression Models	3-features	MSE	4-features	MSE
KNN	SAP_pos_CDRH3 SCM_pos_CDR SCM_neg_CDR	4.711	SAP_pos_CDRH1 SAP_pos_CDRH3 SCM_neg_CDRH3 SCM_pos_CDR	4.579
LR	SCM_neg_CDR SCM_pos_CDRL2 SCM_pos_CDR	5.937	SCM_neg_CDRH1 SCM_neg_CDRL3 SCM_pos_Lv SCM_pos_CDR	5.058
RF	SAP_pos_CDRH2 SAP_pos_CDRH3 SCM_pos_CDR	5.377	SAP_pos_CDRH2 SAP_pos_CDRH3 SCM_neg_Fv SCM_pos_CDR	5.316
SVR	SAP_pos_CDRH3 SCM_pos_CDR SCM_neg_CDRL1	5.884	SAP_pos_CDRH3 SCM_neg_CDR SCM_neg_CDRL1 SCM_pos_CDR	5.958

[a]The feature selection results for other physical properties and the MD are shown in Tables S1−23.

Machine Learning Models for Biophysical Properties Prediction. In this study, we employed machine learning protocols sourced from the scikit-learn library,[39] encompassing linear regression *(linear_model.LinearRegression())*, k-nearest neighbors regression *(neighbors.KNeighborsRegressor())*, support vector regression *(svm.SVR())*, and random forest regression *(ensemble.RandonForestRegressor())* models for our machine learning training.

To optimize the performance of machine learning models, we conducted parameter tuning for each regression algorithm and selected the best parameters based on correlation coefficient values (r) and MSE values. For k-nearest neighbors (KNN) regression, we varied the number of neighbors from $n = 2$ to $n = 8$. For support vector regression (SVR), we explored a range of parameters, adjusting C from 0.1 to 15.0 and ϵ from 0.1 to 10.0. Additionally, for random forest (RF) regression, we adjusted the maximum depth from 2 to 6. In addition to parameter tuning, we considered the number of features used for machine learning training. Initial testing revealed that using only 1 or 2 features did not yield satisfactory model performance. Therefore, we utilized a minimum of 3 features and up to 5 features to train the models, balancing model performance with training time.

Due to the limited data set, we employed the Leave-One-Out-Cross-Validation (LOOCV) method, a commonly used technique for assessing model reliability in such circumstances.[40] With 137 data sets available, each iteration of LOOCV involved training machine learning models on 136 data sets and using the remaining 1 data set for validation. This process was repeated 137 times to obtain 137 prediction values. In our analysis, we anticipated that the correlation coefficient and MSE from LOOCV would likely decrease compared to machine learning models built from all data sets. To assess the reliability of our machine learning models, we established a threshold. If the difference between the coefficient of determination (R^2) of the training and LOOCV was less than 0.3, we deemed the models to be reliable.

RESULTS

Correlation of the Features from MD Simulations and DeepSP. In this research, we used a total of 30 antibody-specific surface descriptors as features: SAP, SCM_neg, and SCM_pos. These properties were calculated across 10 mAb domains, including CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, CDRL3, all CDR, Hv, Lv, and Fv. We calculated these features using MD simulations and the DeepSP surrogate model for the 137 mAbs. In Figure 2, the heatmap depicts the correlation of SCM_pos, SCM_neg, and SAP across antibody domains from DeepSP predictions.

First, comparing the results from MD simulations (Figure 2A) and DeepSP (Figure 2B), we observe a similar distribution of correlation coefficients in both heatmaps, indicating that DeepSP predictions are consistent with MD-derived features. In Figure S2, we note that the worst correlation is SAP_pos_Fv, with a correlation coefficient of 0.6, while the remaining predictions exhibit correlation coefficients of 0.7 or higher (21 out of 30 properties ≥0.85). The calculation time for DeepSP is much less (a few seconds for all 137 mAbs).

For the result from MD simulations, Figure 2A shows that the correlations of SAP in each domain with SCM_pos and SCM_neg range from −0.21 to 0.31; for the result from DeepSP which shown in Figure 2B, the correlations of SAP in each domain with SCM_pos and SCM_neg range from −0.24 to 0.44. The results from MD simulations and DeepSP both indicate a weak correlation. Moreover, we also observe a general negative correlation between SCM_pos and SCM_neg. Regarding the correlation of SAP in individual domains, we find that the SAP_pos_CDRH3 exhibits relatively high correlations with SAP_pos_CDR, SAP_pos_Hv, and SAP_pos_Fv, with correlations of 0.66, 0.78, and 0.52, respectively. For the result from DeepSP, the correlation coefficients are 0.65, 0.70, and 0.51, respectively. Regarding the correlation of SCM_neg in individual domains, we find that the correlation of CDRH3 has a weak correlation with other individual CDR domains. For example, the correlation between SCM_neg_CDRH3 and SCM_neg_CDRH1 and the correlation between SCM_neg_CDRH3 and SCM_neg_CDRH2 is from 0.12 to 0.26 for MD-based features (around 0.2 for DeepSP-based features). SCM_neg_CDR exhibits higher correlations with other SCM_neg regions, ranging from 0.40 to 0.93 for MD-based features (from 0.46 to 0.95 for DeepSP). On the other hand, SCM_neg_Hv shows weak correlations with regions located in the light chain of the mAb. Similar patterns

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

can also be observed in SCM_pos. The diversity of the features that capture different physical properties is advantageous for machine learning models because excessively similar features can impede the effectiveness of model training.[41]

Feature Selection for Machine Learning. With a limited data set, using all 30 features to train machine-learning models can easily lead to overfitting. Therefore, we determined the optimal number of features for machine learning model training using exhaustive feature selection. From preliminary testing, we could not establish predictive models with good performance using fewer than 2-feature combinations. However, extracting 5 features exhaustively would result in as many as 142506 combinations, significantly increasing computation time for feature selection. Therefore, we used at least 3-feature and at most 5-feature combinations for subsequent machine learning model training. In Table 1, using ELISA as an example, we calculated MSE using 4-fold cross-validation under different regression algorithms and feature combinations. The combinations with the lowest MSE values were chosen for subsequent machine learning model training. For instance, the KNN regression model yielded the lowest MSE (4.711) using the 3-feature combination of SAP_pos_CDRH3, SCM_pos_CDR, and SCM_neg_CDR and yielded the lowest MSE (4.579) using the 4-feature combination of SAP_pos_CDRH1, SAP_pos_CDRH3, SCM_neg_CDRH3, and SCM_pos_CDR.

Machine Learning Models for 12 Biophysical Properties Prediction. We identified the best feature combinations from the feature selection process for building predictive models of 12 biophysical properties. The features were selected from the DeepSP method and MD simulations. Hence, in this study, a total of 24 machine-learning models were evaluated for their performance. Among them, ELISA prediction showed the best result, with the MD model exhibiting a correlation coefficient (r) of 0.89 and MSE of 1.50, while the DeepSP model achieved an r of 0.85 and MSE of 1.97 (Figure 3). Furthermore, Figure 3 includes the results of Leave-One-Out-Cross-Validation (LOOCV), a valuable tool for assessing model performance in small data sets. The MD model showed an r of 0.73 and MSE of 3.26, whereas the DeepSP model had an r of 0.66 and MSE of 3.98. For the ELISA predictive model, the best-performing model was achieved using the KNN regression algorithm with a parameter k-nearest neighbors value of k = 3. This model utilized 3 features, SAP_pos_CDRH3, SCM_pos_CDR, and SCM_neg_CDR, extracted using the exhaustive selection method. Comparing the MSE before and after parameter tuning, we observed a significant reduction in MSE values. In Figure 3 and Table 1, the MSE value decreased from 4.711 to 1.50 (for MD_all), 3.26 (for MD_LOOCV), 1.97 (for DeepSP_all), and 3.98 (for DeepSP_LOOCV) after parameter tuning. The performance of the remaining 11 models is depicted in Figure S3 (DeepSP) and Figure S4 (MD simulations).

Comparing Table 2 with Tables S1−11, S24, and S12−23, we noticed that for both the MD and DeepSP features, the MSE values decreased after tuning the parameters with the selected features, except for the predictive models of AS. For the predictive models of AS, we encountered challenges in establishing reliable model performance. In the result from DeepSP, the highest correlation coefficient obtained was 0.27, while in the LOOCV, the correlation coefficient was 0.08 (Figure S3). For MD-based features, the best model achieved a correlation of 0.21.

In evaluating the performance of the machine learning models across the 12 biophysical properties, we observed that while the

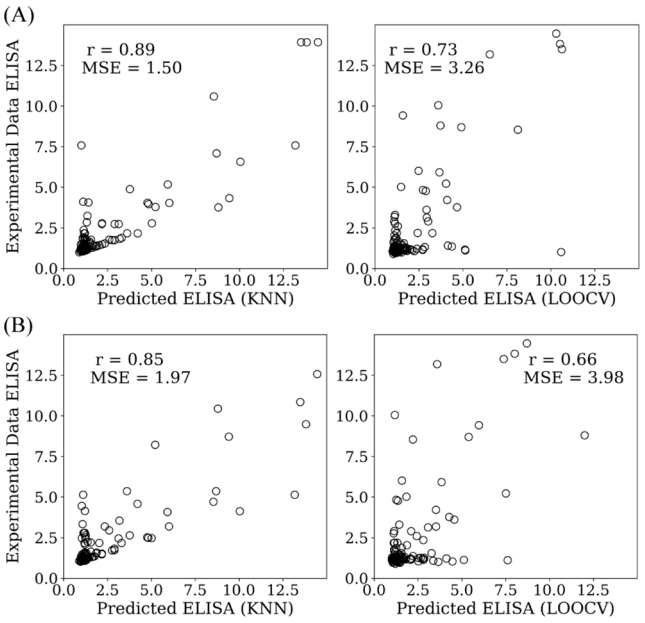


Figure 3. Correlation coefficients for the 3-feature k-nearest neighbors (KNN) models trained using the entire data set of 137 samples with LOOCV for ELISA predictions. (A) the machine learning model training using features extracted from MD simulations; (B) the machine learning model training using features extracted from DeepSP. The hyperparameters of this model were set to k = 3.

models performed well for several properties (e.g., ELISA, HIC, BVP), their predictive performance for AS was notably lower. Specifically, the correlation coefficient for AS remained low in both the MD and DeepSP models, and the models struggled to capture meaningful patterns. Upon further examination, we hypothesize that the weaker performance for AS may be attributable to the unique characteristics of the assay and the complexity of the biophysical mechanisms it measures. AS measures the stability of monoclonal antibodies under accelerated conditions, often capturing subtle and multifactorial changes in conformation and aggregation propensity that may not be fully captured by the surface properties (such as hydrophobicity and charge) used as features in our models. Furthermore, the AS assay likely involves factors beyond the direct molecular surface descriptors used in the models, such as internal protein conformational stability, disulfide bond stability, or solvent interactions under stress conditions. These factors may contribute to AS results not following the trends observed for other biophysical properties, such as HIC or SGAC, which are more directly influenced by surface charge and hydrophobicity. This suggests that additional features, possibly capturing intramolecular interactions or protein flexibility, could be required to improve the model's ability to predict AS with greater accuracy.

Additionally, even in the LOOCV, positive *r* values were not attained (Figure S4). The 12 models represent various properties that have different units and scales. Therefore, we normalized the MSE values into the Normalized Mean Squared Error (NMSE) and summarized them in Table 2. We found that better model performance, which has a higher correlation, aligns with lower NMSE. In Table 3, we summarize the performance of machine learning models for the 12 properties, using MD and DeepSP features, along with the model developed by Hebditch and Warwicker.[28] The MD or DeepSP models provide better predictions for protein binding and hydrophobicity properties,

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

Table 2. Performance Comparison of Machine Learning Models and Their LOOCV for Predicting 12 Biophysical Properties of mAbs using Different Feature Sets from DeepSP[a,b]

Properties	Best model performance	Feature	MSE/NMSE (All data)	MSE/NMSE (LOOCV)
ACSINS	SVR	SAP_pos_CDRH1 SAP_pos_CDRL3 SCM_pos_CDRH1 SCM_neg_CDR	$0.690/4.16 \times 10^{-2}$	$1.00/6.03 \times 10^{-2}$
AS	LR	SAP_pos_CDRH2 SCM_pos_CDRL2 SCM_pos_CDRL3 SCM_neg_CDRL3	$0.003/3.16 \times 10^{-2}$	$0.003/3.16 \times 10^{-2}$
BVP	KNN	SAP_pos_CDRH1 SAP_pos_CDRH3 SCM_pos_CDR SCM_neg_CDRH3	$6.43/1.36 \times 10^{-2}$	$11.95/2.53 \times 10^{-2}$
CIC	KNN	SAP_pos_CDRL2 SAP_pos_CDRL3 SAP_pos_Lv SCM_neg_CDR	$0.570/2.88 \times 10^{-2}$	$0.80/4.04 \times 10^{-2}$
CSI	SVR	SAP_pos_CDRL1 SAP_pos_Lv SCM_pos_CDRH2 SCM_neg_CDRL2	$0.88/4.81 \times 10^{-2}$	$1.08/5.90 \times 10^{-2}$
ELISA	KNN	SAP_pos_CDRH3 SCM_pos_CDR SCM_neg_CDR	$1.97/1.07 \times 10^{-2}$	$3.98/2.16 \times 10^{-2}$
HIC	SVR	SAP_pos_CDRL3 SAP_pos_CDR SAP_pos_Hv SCM_pos_CDRH3	$0.56/2.06 \times 10^{-2}$	$0.81/2.98 \times 10^{-2}$
HEK	KNN	SAP_pos_CDRH2 SAP_pos_CDRL3 SCM_pos_Lv SCM_neg_Lv	$2236.53/3.05 \times 10^{-2}$	$3367.50/4.60 \times 10^{-2}$
PSR	SVR	SAP_pos_Lv SCM_pos_CDRH2 SCM_neg_CDRL2	$0.02/3.05 \times 10^{-2}$	$0.03/4.57 \times 10^{-2}$
SGAC	SVR	SAP_pos_CDRH1 SAP_pos_CDRL3 SCM_neg_CDRH2 SCM_neg_Lv	$0.52/5.30 \times 10^{-2}$	$0.82/8.36 \times 10^{-2}$
SMAC	KNN	SAP_pos_CDR SAP_pos_Fv SCM_neg_CDRL2 SCM_neg_Fv	$0.61/2.58 \times 10^{-2}$	$0.87/3.68 \times 10^{-2}$
Tm	KNN	SAP_pos_CDRH1 SAP_pos_CDRH2 SCM_pos_CDRH3	$20.78/2.03 \times 10^{-2}$	$29.04/2.84 \times 10^{-2}$

[a]The feature sets from MD simulations are shown in Table S24. [b]NMSE: Normalized Mean Squared Error.

such as ELISA, HIC, and BVP. This is closely related to the features we used because SCM and SAP represent the surface charge and hydrophobic distribution of mAbs, which in turn affect their hydrophobicity and aggregation propensity.[42,43] On the other hand, some properties related to AS and Tm, which are related to conformational stability, cannot be predicted accurately. In Jain et al. study, Tm was measured using differential scanning fluorimetry (DSF), which monitors thermal unfolding as hydrophobic regions are exposed and bind to fluorescent dyes. Therefore, Tm reflects the thermal unfolding mechanism leading to conformational changes and stability. Given SAP and SCM features primarily describe surface properties, they may not be ideal for predicting properties related to conformational stability. Furthermore, we compared our models with those of Hebditch and Warwicker.[28] In Hebditch's work, amino acid sequence-based features were utilized to construct 12 ML models for the prediction of biophysical properties. Their methodology includes experimental data transformation and subsequent feature selections based on the Variance Inflation Factor (VIF). Their study employed cross-validation (50 times repeated 10-fold) to estimate the models' performance on unseen data. Both our work and Hebditch's study use the same 137 mAbs data set, which was originally provided in the Jain et al. study. Generally, the

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

Table 3. Summary of Machine Learning Results (R^2) for 12 Physical Properties Prediction[a,b]

MD	HIC	SMAC*	CIC	ACSINS*	ELISA	BVP	SGAC*	PSR	HEK	Tm	CSI*	AS
Algorithm	KNN	SVR	SVR	SVR	KNN	SVR	KNN	SVR	RF	SVR	SVR	LR
R^2(Model)	0.640	0.578	0.462	0.490	0.792	0.656	0.476	0.533	0.476	0.194	0.449	0.044
R^2(LOOCV)	0.384	0.348	0.303	0.212	0.533	0.436	0.194	0.314	0.185	0.084	0.203	−0.05
DeepSP												
Algorithm	SVR	KNN	KNN	SVR	KNN	KNN	SVR	SVR	KNN	KNN	SVR	LR
R^2(Model)	0.533	0.548	0.476	0.464	0.723	0.548	0.476	0.476	0.423	0.397	0.314	0.073
R^2(LOOCV)	0.325	0.348	0.260	0.230	0.436	0.360	0.176	0.292	0.137	0.160	0.160	0.006
Hebditch group[28]												
Algorithm	EN	EN	SVR	EN	RF	RF	SVR	SVR	SVR	SVR	SVR	SVR
R^2(Model)	0.391	0.353	0.306	0.268	0.383	0.355	0.215	0.316	0.112	0.130	0.169	0.086

[a]KNN: k-nearest neighbor; SVR: support vector regression; RF: random forest; LR: linear regression; EN: elastic net. [b]The asterisk represents that the models are trained by transformed experimental data.

performance of these models is good, though some variability is noted across different biophysical techniques. The performance metrics provided in the study are based on the entire data set rather than a separate validation set, as there is no explicit mention of a validation group in their results tables. This suggests that the reported model performance metrics are likely for the overall model performance on the complete data set. In our research, we find that, whether using features from MD simulations or from DeepSP, the machine learning models trained on the entire data set show performance that, except for AS prediction, have R^2 values exceeding those of Hebditch's models. For example, the R^2 values for the HIC models were 0.640 (MD) and 0.533 (DeepSP), respectively. The HIC model from Hebditch et al, using the whole data to train, has an R^2 value of 0.391. We observed that even our validation test for several properties performed better than Hebditch's models, which used the entire data set for training.

Mechanistic Insight of the Machine Learning Features. Understanding the correlation between selected features and these 12 biophysical properties can enhance our comprehension of their physical significance. In Figure 4, the heatmaps display the correlation coefficients between each feature, derived from (A) MD simulations and (B) the DeepSP, and the 12 properties. It is evident from these figures that the features from MD simulations and results from DeepSP are similar, showing a comparable pattern in the correlation coefficients. This similarity suggests that the DeepSP model can predict results similar to those obtained from MD simulations.

Protein aggregation propensity is driven by hydrophobicity; therefore, for hydrophobicity properties (HIC, SMAC, SGAC), numerous SAP features were selected from MD and DeepSP features. It is observed that domains in the heavy chain have high correlation coefficients with these three hydrophobicity-related properties. The highest is MD-derived SAP_pos_CDR (0.38), although it was not used in the subsequent MD-based machine learning model building. This is because other feature combinations of the SAP domain achieved better results in our feature selection methodology, and the combination performs better than a single feature.

The correlations of SGAC, SMAC, and HIC with the SAP features are closely aligned. However, the SGAC data exhibit a reversed correlation pattern relative to HIC and SMAC. Lower SGAC indicates increased hydrophobicity, contrasting with the positive correlations observed in HIC and SMAC, where higher values indicate greater hydrophobicity.

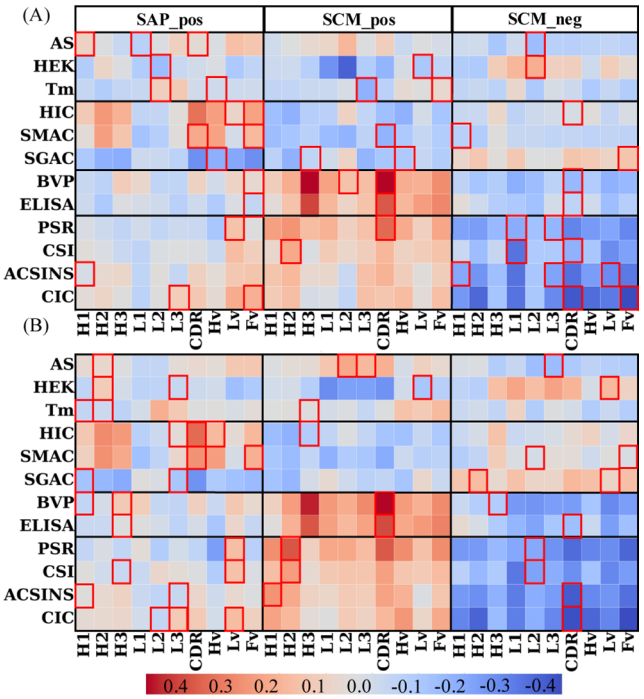


Figure 4. Correlation heatmap of the features from (A) MD simulations and (B) DeepSP across antibody domains with 12 biophysical properties. Red boxes represent the feature combinations for the best machine-learning models for each property. H1 refers to CDRH1, H2 to CDRH2, H3 to CDRH3, L1 to CDRL1, L2 to CDRL2, and L3 to CDRL3. The correlation values represented in the heatmaps are provided in the supplementary file: HeatMap_SI.xlsx.

Regarding binding-related properties (ELISA, BVP), these properties also demonstrated the best performance in predictive model development in this study. For BVP models, regardless of whether the features are from MD simulations or DeepSP, SCM_pos_CDRH3 and SCM_pos_CDR both show high correlation coefficients with BVP, at 0.54 (MD simulations) and 0.47 (DeepSP), and 0.54 (MD simulations) and 0.49 (DeepSP), respectively. Although both SCM_pos_CDRH3 and SCM_pos_CDR show a high correlation with BVP, only SCM_pos_CDR was ultimately selected because CDRH3 is included within CDR, thus BVP models only use SCM_pos_CDR combined with other features with lower correlation coefficients to further improve the performance. A similar scenario occurred in the building of ELISA predictive models.

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

In terms of self/cross-interaction related properties (ACSINS, CSI, PSR, CIC), we find that SCM_pos and SCM_neg exhibit strongly positive or negative correlations with these properties, respectively. In contrast, the correlation of SAP features with these properties does not exhibit distinct patterns compared with SCM_pos or SCM_neg. Charge distribution often has a more significant impact on mAb's self-and cross interactions than hydrophobicity, as charge directly influences the electrostatic forces between molecules, which play a critical role in biomolecular interactions.[44,45] Dividing these four properties into self-interaction (ACSINS, CSI) and cross-interaction groups (PSR, CIC), it is evident that self-interaction-related properties are more influenced by SCM_neg, with high negative correlations such as −0.34 (MD simulations), and −0.38 (DeepSP) for ACSINS, and −0.33 (MD simulations), and −0.32 (DeepSP) for CSI. The increased surface negative charge on mAbs reduces self-interaction due to repulsive forces. For cross-interaction-related properties, SCM_pos has a more substantial influence, showing that the strongest correlations are the positive correlation of 0.40 (MD simulations) and 0.39 (DeepSP) for PSR, respectively.

Cross-interaction assay, including ELISA, BVP, CIC, PSR, measured the potential for antibodies to bind nonspecifically to off-target molecules. As noted in a recent review, nonspecific interactions are often driven by surface patches—clusters of exposed amino acid residues with similar physicochemical properties, such as positive charge, which promote nonspecific binding to negatively charged targets.[46] The overall positive charge correlates with nonspecific binding in these assays due to the attraction of the positively charged surface patches to negatively charged ligands. In contrast, self-interaction assays, such as CSI and ACSINS, capture interactions between identical antibody molecules. These interactions are influenced by both repulsive and attractive electrostatic forces. Strong positive or negative charges on the surface can either enhance repulsion or encourage binding, leading to the observed strong absolute charge correlations for self-interactions. This aligns with the understanding that uneven charge distributions, such as charged surface patches, can promote self-association and nonspecific binding.

Lastly, regarding properties like Tm, HEK, and AS, it is challenging to find a pattern among our current features, as these properties are more closely related to conformational stability and intramolecular interactions, not intermolecular interactions. Tm is affected by intramolecular hydrogen bonds, disulfide bonds, and ionic bonds.[47−49] For HEK titer, expression levels are influenced by variations in certain amino acid sequences, particularly signal sequences and N-terminal sequences.[50,51] AS, a key assay for assessing the long-term aggregation tendencies of mAbs, effectively evaluates their stability. Importantly, the amino acid composition and the interactions between amino acids of mAbs significantly influence the results of this assay.[52−55] Additionally, in Jain's study, a low concentration condition (1 mg/mL) was used for accelerated stability testing. However, at such a low concentration, the mAb may not exhibit the full range of stability behaviors seen at higher concentrations, potentially leading to different aggregation or degradation profiles. The lack of interactions at low concentrations may result in the poor prediction of our models. These specific sequences can greatly improve the flexibility and stability of mAbs, impacting their overall performance in the AS test. The features used in this work describe the surface hydrophobicity and charge distribution, which shows better prediction for intermolecular interactions. Future work will include more features to describe intramolecular interactions and conformational stability.

Web Application Development and Models Code Source. We have integrated the predictive models for these 12 biophysical properties, based on features predicted by DeepSP, into a web application, AbDev, accessible at https://devpred.onrender.com/AbDev. This tool is provided for use in mAb developability research. Users simply need to input the sequences of the light and heavy chain variable regions of a mAb. AbDev will then provide the predicted values for these 12 biophysical properties. Moreover, since the experimental data for 5 of the 12 properties—ACSINS, CSI, CIC, SGAC, and SMAC—underwent data transformation, the resultant predictions can be challenging to interpret directly. Consequently, we employ threshold values for 10 out of the 12 experimental platforms (excluding Tm and HEK) derived from the original study by Jain et al.[25] The threshold values for these 5 properties are also transformed in the same manner and set as new threshold values. Within the webapp, once the predicted results are obtained, any values exceeding these thresholds will be highlighted, indicating that the values do not meet the standard and pose potential issues for the developability of the mAbs. Furthermore, for those planning to process a large number of mAb sequences at once, we also provide codes and scripts that can be accessed via https://github.com/Lailabcode/AbDev.

DISCUSSION

In this study, we constructed several machine learning models to predict 12 biophysical properties using molecular modeling and dynamics simulation. However, this process is time-consuming and demands significant computational resources. The simulation of a single mAb sample required approximately 20−24 h using one NVIDIA V100 GPU. Moreover, MD simulation results can vary due to molecular fluctuations, leading to inconsistent outcomes. To address these challenges, our group recently developed a deep learning model called DeepSP, which rapidly predicts the features needed for machine learning models by providing the sequence of the Fv regions. In previous research, we demonstrated that the SCM and SAP values obtained from DeepSP could effectively build machine-learning models for predicting the high-concentration aggregation rate of mAbs. Since the aggregation rate of mAbs is closely linked to protein interactions arising from hydrophobicity and charge distribution, and some of the 12 biophysical properties under study are related to protein hydrophobic and charge interactions, we hypothesize that using SCM and SAP as features have the potential to construct predictive models for these properties as well.

In terms of results, while the model performance using features from DeepSP is slightly inferior to those using MD features, it still produces acceptable outcomes. It is also observed that the best feature combinations for training machine learning models do not always include the single feature with the highest correlation. Often, features that are highly correlated with the target can also be highly correlated with each other, potentially leading to multicollinearity, which can adversely affect the models' performance. This suggests that relying solely on the highest correlation features is unnecessary; various feature combinations can also successfully build models with robust performance. Moreover, in our analysis of the models that predicted poorly, properties such as AS, HEK, and Tm displayed weak correlations across all features. However, we successfully

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153

developed predictive models with strong performance, particularly regarding mAbs binding affinity's biophysical properties like ELISA and BVP. Predictive models concerning mAbs self/cross-interaction also performed well. This indicates that SAP and SCM scores are suitable protein descriptors for these three properties. Additionally, all 137 mAbs included in the data set are clinical trial candidates and have been meticulously designed to the best of our knowledge. This implies that these models lack input data for poorly designed mAbs, which is one of the reasons for the limited performance observed for some properties.

CONCLUSION

In this study, we have successfully constructed and evaluated multiple machine learning models to predict a set of 12 critical biophysical properties of mAbs using data obtained from both MD simulations and a novel deep learning approach, DeepSP. The incorporation of DeepSP has significantly enhanced our methodology, providing a faster and equally effective alternative to MD simulations for extracting essential features required for model training, thereby addressing the significant computational time and resource constraints associated with MD simulations. Our results demonstrate that the predictive models developed herein significantly outperform existing methods, particularly in predicting properties related to the intermolecular interactions of mAbs, which are critical for their developability as therapeutic agents. The successful application of machine learning, especially the DeepSP model, highlights the potential for these computational approaches to accelerate the development process of mAbs by enabling rapid and accurate prediction of SAP and SCM as the features. Moreover, the methodology outlined in our study provides a framework for employing computational tools in the early stages of mAb development, thereby potentially reducing the time and cost associated with experimental testing. It is evident that integrating machine learning and deep learning techniques into the mAb development process presents a promising avenue for enhancing the efficiency and success rate of therapeutic antibody discovery and development.

Despite the successes, certain limitations, such as the predictive accuracy for properties related to conformational and thermal stability, need to be addressed in future work. Additionally, our findings underscore the importance of feature selection and the necessity to balance model complexity and predictive power to avoid overfitting, especially when dealing with limited data sets. In conclusion, this study represents a significant step forward in the application of machine learning and deep learning methods for predicting mAb biophysical properties. Our models provide better predictive performance compared to previous approaches and offer a more efficient and cost-effective solution for the early stage screening of mAbs, thereby facilitating the development of more effective and safer therapeutic agents.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.4c00804.

Figure S1 shows the experimental data distribution after transformation, which can help visualize how the data normalization or transformation was applied to prepare it for more effective machine learning analysis; Figure S2 compares the SAP and SCM values calculated from MD simulations and those calculated from DeepSP; Figures S3 and S4 summarize the performance of the 12 machine learning models trained from features predicted by DeepSP and MD simulations, respectively; Tables S1−23 list the MSE values obtained before parameter tuning, using different feature combinations and regression algorithms. Table S24 lists the MSE and NMSE values for the 12 models trained with MD simulations-based features after parameter tuning (PDF)

Numerical values of the heatmap of Figures 2 and 4 (XLSX)

AUTHOR INFORMATION

Corresponding Author

Pin-Kuang Lai − *Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken 07030, New Jersey;* orcid.org/0000-0003-2894-3900; Email: plai3@stevens.edu

Authors

I-En Wu − *Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken 07030, New Jersey*

Lateefat Kalejaye − *Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken 07030, New Jersey*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.molpharmaceut.4c00804

Author Contributions

I.-E.W.: Writing−review and editing, Writing−original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. L.K.: Methodology, Data curation. P.-K.L.: Writing−review and editing, Supervision, Methodology, Funding acquisition, Conceptualization

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the Biomolecular Interaction Technologies Center at the University of Delaware (UDR0000314). We thank the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support for supporting computing resources.

REFERENCES

(1) Walsh, G. Biopharmaceutical Benchmarks 2018. *Nat. Biotechnol.* **2018**, *36* (12), 1136−1145.

(2) Anselmo, A. C.; Gokarn, Y.; Mitragotri, S. Non-Invasive Delivery Strategies for Biologics. *Nat. Rev. Drug Discovery* **2019**, *18* (1), 19−40.

(3) El Abd, Y.; Tabll, A.; Smolic, R.; Smolic, M. Mini-Review: The Market Growth of Diagnostic and Therapeutic Monoclonal Antibodies - SARS CoV-2 as an Example. *Hum. Antibodies* **2022**, *30* (1), 15−24.

(4) Ranjan, S.; Chung, W. K.; Zhu, M.; Robbins, D.; Cramer, S. M. Implementation of an Experimental and Computational Tool Set to Study Protein-mAb Interactions. *Biotechnol. Prog.* **2019**, *35* (4), No. e2825.

(5) Kim, J.; McFee, M.; Fang, Q.; Abdin, O.; Kim, P. M. Computational and Artificial Intelligence-Based Methods for Antibody Development. *Trends Pharmacol. Sci.* **2023**, *44* (3), 175−189.

(6) Khetan, R.; Curtis, R.; Deane, C. M.; Hadsund, J. T.; Kar, U.; Krawczyk, K.; Kuroda, D.; Robinson, S. A.; Sormanni, P.; Tsumoto, K.; Warwicker, J.; Martin, A. C. R. Current Advances in Biopharmaceutical

Informatics: Guidelines, Impact and Challenges in the Computational Developability Assessment of Antibody Therapeutics. *mAbs* **2022**, *14* (1), 2020082.
(7) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9* (1), 71.
(8) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035−4061.
(9) Lapelosa, M.; Patapoff, T. W.; Zarraga, I. E. Molecular Simulations of the Pairwise Interaction of Monoclonal Antibodies. *J. Phys. Chem. B* **2014**, *118* (46), 13132−13141.
(10) Al Qaraghuli, M. M.; Kubiak-Ossowska, K.; Mulheran, P. A. Thinking Outside the Laboratory: Analyses of Antibody Structure and Dynamics within Different Solvent Environments in Molecular Dynamics (MD) Simulations. *Antibodies* **2018**, *7* (3), 21.
(11) Brandt, J. P.; Patapoff, T. W.; Aragon, S. R. Construction, MD Simulation, and Hydrodynamic Validation of an All-Atom Model of a Monoclonal IgG Antibody. *Biophys. J.* **2010**, *99* (3), 905−913.
(12) Zamolo, L.; Busini, V.; Moiani, D.; Moscatelli, D.; Cavallotti, C. Molecular Dynamic Investigation of the Interaction of Supported Affinity Ligands with Monoclonal Antibodies. *Biotechnol. Prog.* **2008**, *24* (3), 527−539.
(13) Fernández-Quintero, M. L.; Kokot, J.; Waibl, F.; Fischer, A.-L. M.; Quoika, P. K.; Deane, C. M.; Liedl, K. R. Challenges in Antibody Structure Prediction. *mAbs* **2023**, *15* (1), 2175319.
(14) Leem, J.; Dunbar, J.; Georges, G.; Shi, J.; Deane, C. M. ABodyBuilder: Automated Antibody Structure Prediction with Data-Driven Accuracy Estimation. *mAbs* **2016**, *8* (7), 1259−1268.
(15) Phillips, A.; Srinivas, A.; Prentoska, I.; O'Dea, M.; Kustrup, M.; Hurley, S.; Bruno, S.; Nguyen, V.; Lai, P.-K. Teaching Biologics Design Using Molecular Modeling and Simulations. *Biochem. Mol. Biol. Educ.* **2024**, *52* (3), 299−310.
(16) Mason, D. M.; Friedensohn, S.; Weber, C. R.; Jordi, C.; Wagner, B.; Meng, S. M.; Ehling, R. A.; Bonati, L.; Dahinden, J.; Gainza, P.; Correia, B. E.; Reddy, S. T. Optimization of Therapeutic Antibodies by Predicting Antigen Specificity from Antibody Sequence via Deep Learning. *Nat. Biomed. Eng.* **2021**, *5* (6), 600−612.
(17) Ali, S. A.; Hassan, M. I.; Islam, A.; Ahmad, F. A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Curr. Protein Pept. Sci.* **2014**, *15* (5), 456−476.
(18) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Design of Therapeutic Proteins with Enhanced Stability. *Proc. Natl. Acad. Sci.* **2009**, *106* (29), 11937−11942.
(19) Agrawal, N. J.; Helk, B.; Kumar, S.; Mody, N.; Sathish, H. A.; Samra, H. S.; Buck, P. M.; Li, L.; Trout, B. L. Computational Tool for the Early Screening of Monoclonal Antibodies for Their Viscosities. *mAbs* **2016**, *8* (1), 43−48.
(20) Lai, P.-K.; Gallegos, A.; Mody, N.; Sathish, H. A.; Trout, B. L. Machine Learning Prediction of Antibody Aggregation and Viscosity for High Concentration Formulation Development of Protein Therapeutics. *mAbs* **2022**, *14* (1), 2026208.
(21) Lai, P.-K.; Fernando, A.; Cloutier, T. K.; Kingsbury, J. S.; Gokarn, Y.; Halloran, K. T.; Calero-Rubio, C.; Trout, B. L. Machine Learning Feature Selection for Predicting High Concentration Therapeutic Antibody Aggregation. *J. Pharm. Sci.* **2021**, *110* (4), 1583−1591.
(22) Raybould, M. I. J.; Marks, C.; Krawczyk, K.; Taddese, B.; Nowak, J.; Lewis, A. P.; Bujotzek, A.; Shi, J.; Deane, C. M. Five Computational Developability Guidelines for Therapeutic Antibody Profiling. *Proc. Natl. Acad. Sci.* **2019**, *116* (10), 4025−4030.
(23) Vora, L. K.; Gholap, A. D.; Jetha, K.; Thakur, R. R. S.; Solanki, H. K.; Chavda, V. P. Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics* **2023**, *15* (7), 1916.
(24) Habiballah, S.; Reisfeld, B. Adapting Physiologically-Based Pharmacokinetic Models for Machine Learning Applications. *Sci. Rep.* **2023**, *13* (1), 14934.
(25) Jain, T.; Sun, T.; Durand, S.; Hall, A.; Houston, N. R.; Nett, J. H.; Sharkey, B.; Bobrowicz, B.; Caffry, I.; Yu, Y.; Cao, Y.; Lynaugh, H.; Brown, M.; Baruah, H.; Gray, L. T.; Krauland, E. M.; Xu, Y.; Vásquez, M.; Wittrup, K. D. Biophysical Properties of the Clinical-Stage Antibody Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (5), 944−949.
(26) Dzisoo, A. M.; Kang, J.; Yao, P.; Klugah-Brown, B.; Mengesha, B. A.; Huang, J. SSH: A Tool for Predicting Hydrophobic Interaction of Monoclonal Antibodies Using Sequences. *BioMed. Res. Int.* **2020**, *2020*, 3508107.
(27) Waibl, F.; Fernández-Quintero, M. L.; Wedl, F. S.; Kettenberger, H.; Georges, G.; Liedl, K. R. Comparison of Hydrophobicity Scales for Predicting Biophysical Properties of Antibodies. *Front. Mol. Biosci.* **2022**, *9*, 960194.
(28) 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*, No. e8199.
(29) 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.
(30) Dietterich, T. Overfitting and Undercomputing in Machine Learning. *ACM Comput. Surv.* **1995**, *27* (3), 326−327.
(31) Padlan, E. A. Anatomy of the Antibody Molecule. *Mol. Immunol.* **1994**, *31* (3), 169−217.
(32) Boehm, M. K.; Woof, J. M.; Kerr, M. A.; Perkins, S. J. The Fab and Fc Fragments of IgA1 Exhibit a Different Arrangement from That in IgG: A Study by X-Ray and Neutron Solution Scattering and Homology Modelling1. *J. Mol. Biol.* **1999**, *286* (5), 1421−1447.
(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926−935.
(34) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.
(35) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D., Jr.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B* **2010**, *114* (23), 7830−7843.
(36) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781−1802.
(37) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71−73.
(38) Raschka, S. MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack. *J. Open Source Softw.* **2018**, *3* (24), 638.
(39) 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.
(40) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data mining, inference, and prediction*; Springer Series in Statistics: New York, NY, 2009. .
(41) Miller, N. L.; Clark, T.; Raman, R.; Sasisekharan, R. Learned Features of Antibody-Antigen Binding Affinity. *Front. Mol. Biosci.* **2023**, *10*, 1112738.
(42) Fekete, S.; Veuthey, J.-L.; Beck, A.; Guillarme, D. Hydrophobic Interaction Chromatography for the Characterization of Monoclonal Antibodies and Related Products. *J. Pharm. Biomed. Anal.* **2016**, *130*, 3−18.
(43) Yadav, S.; Laue, T. M.; Kalonia, D. S.; Singh, S. N.; Shire, S. J. The Influence of Charge Distribution on Self-Association and Viscosity Behavior of Monoclonal Antibody Solutions. *Mol. Pharmaceutics* **2012**, *9* (4), 791−802.
(44) Roberts, D.; Keeling, R.; Tracka, M.; van der Walle, C. F.; Uddin, S.; Warwicker, J.; Curtis, R. The Role of Electrostatics in Protein−Protein Interactions of a Monoclonal Antibody. *Mol. Pharmaceutics* **2014**, *11* (7), 2475−2489.

(45) Boswell, C. A.; Tesar, D. B.; Mukhyala, K.; Theil, F.-P.; Fielder, P. J.; Khawli, L. A. Effects of Charge on Antibody Tissue Distribution and Pharmacokinetics. *Bioconjugate Chem.* **2010**, *21* (12), 2153−2163.
(46) Ausserwöger, H.; Schneider, M. M.; Herling, T. W.; Arosio, P.; Invernizzi, G.; Knowles, T. P. J.; Lorenzen, N. Non-Specificity as the Sticky Problem in Therapeutic Antibody Development. *Nat. Rev. Chem.* **2022**, *6* (12), 844−861.
(47) Lecerf, M.; Kanyavuz, A.; Lacroix-Desmazes, S.; Dimitrov, J. D. Sequence Features of Variable Region Determining Physicochemical Properties and Polyreactivity of Therapeutic Antibodies. *Mol. Immunol.* **2019**, *112*, 338−346.
(48) Li, W.; Prabakaran, P.; Chen, W.; Zhu, Z.; Feng, Y.; Dimitrov, D. S. Antibody Aggregation: Insights from Sequence and Structure. *Antibodies* **2016**, *5* (3), 19.
(49) Tanghe, M.; Danneels, B.; Last, M.; Beerens, K.; Stals, I.; Desmet, T. Disulfide Bridges as Essential Elements for the Thermostability of Lytic Polysaccharide Monooxygenase LPMO10C from Streptomyces Coelicolor. *Protein Eng., Des. Sel.* **2017**, *30* (5), 401−408.
(50) Güler-Gane, G.; Kidd, S.; Sridharan, S.; Vaughan, T. J.; Wilkinson, T. C. I.; Tigue, N. J. Overcoming the Refractory Expression of Secreted Recombinant Proteins in Mammalian Cells through Modification of the Signal Peptide and Adjacent Amino Acids. *PLoS One* **2016**, *11* (5), No. e0155340.
(51) Powers, J. A.; Skinner, B.; Davis, B. S.; Biggerstaff, B. J.; Robb, L.; Gordon, E.; de Souza, W. M.; Fumagalli, M. J.; Calvert, A. E.; Chang, G.-J. Development of HEK-293 Cell Lines Constitutively Expressing Flaviviral Antigens for Use in Diagnostics. *Microbiol. Spectr.* **2022**, *10* (3), No. e00592−22.
(52) Lu, X.; Nobrega, R. P.; Lynaugh, H.; Jain, T.; Barlow, K.; Boland, T.; Sivasubramanian, A.; Vásquez, M.; Xu, Y. Deamidation and Isomerization Liability Analysis of 131 Clinical-Stage Antibodies. *mAbs* **2019**, *11* (1), 45−57.
(53) Xu, Y.; Wang, D.; Mason, B.; Rossomando, T.; Li, N.; Liu, D.; Cheung, J. K.; Xu, W.; Raghava, S.; Katiyar, A.; et al. Structure, Heterogeneity and Developability Assessment of Therapeutic Antibodies. *mAbs* **2019**, *11* (2), 239−264.
(54) Wang, X.; Singh, S. K.; Kumar, S. Potential Aggregation-Prone Regions in Complementarity-Determining Regions of Antibodies and Their Contribution Towards Antigen Recognition: A Computational Analysis. *Pharm. Res.* **2010**, *27* (8), 1512−1529.
(55) Courtois, F.; Agrawal, N. J.; Lauer, T. M.; Trout, B. L. Rational Design of Therapeutic mAbs against Aggregation through Protein Engineering and Incorporation of Glycosylation Motifs Applied to Bevacizumab. *mAbs* **2016**, *8* (1), 99−112.

https://doi.org/10.1021/acs.molpharmaceut.4c00804
Mol. Pharmaceutics 2025, 22, 142−153