MABS
2022, VOL. 14, NO. 1, e2026208 (12 pages)
https://doi.org/10.1080/19420862.2022.2026208

Taylor & Francis
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics

Pin-Kuang Lai[a,b], Austin Gallegos[c], Neil Mody[c], Hasige A. Sathish[c], and Bernhardt L. Trout[a]

[a]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; [b]Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken, New Jersey, USA; [c]Dosage Form Design and Development, AstraZeneca, Gaithersburg, Maryland, USA

ABSTRACT

Machine learning has been recently used to predict therapeutic antibody aggregation rates and viscosity at high concentrations (150 mg/ml). These works focused on commercially available antibodies, which may have been optimized for stability. In this study, we measured accelerated aggregation rates at 45°C and viscosity at 150 mg/ml for 20 preclinical and clinical-stage antibodies. Features obtained from molecular dynamics simulations of the full-length antibody and sequences were used for machine learning model construction. We found a k-nearest neighbors regression model with two features, spatial positive charge map on the CDRH2 and solvent-accessible surface area of hydrophobic residues on the variable fragment, gives the best performance for predicting antibody aggregation rates (r = 0.89). For the viscosity classification model, the model with the highest accuracy is a logistic regression model with two features, spatial negative charge map on the heavy chain variable region and spatial negative charge map on the light chain variable region. The accuracy and the area under precision recall curve of the classification model from validation tests are 0.86 and 0.70, respectively. In addition, we combined data from another 27 commercial mAbs to develop a viscosity predictive model. The best model is a logistic regression model with two features, number of hydrophobic residues on the light chain variable region and net charges on the light chain variable region. The accuracy and the area under precision recall curve of the classification model are 0.85 and 0.6, respectively. The aggregation rates and viscosity models can be used to predict antibody stability to facilitate pharmaceutical development.

ARTICLE HISTORY
Received 28 October 2021
Revised 19 December 2021
Accepted 4 January 2022

KEYWORDS
Machine learning; molecular dynamics simulations; antibody aggregation; antibody viscosity; developability

Introduction

In recent years, high concentration antibody formulations have been developed for low-volume, subcutaneous administration of therapeutic antibodies and the industry is moving toward convenient, patient-centric dosing schemes that enable at-home delivery.[1] The developability properties of monoclonal antibodies (mAbs), such as low aggregation propensity and low viscosity, are essential to new drug development.[2–4] However, the stability profiles of antibodies at high concentrations are difficult to assess during early-stage discovery and candidate screening due to the limited number of molecules for which sequence, biophysical property data, and sufficient material are available. Therefore, development of predictive tools that can evaluate the developability of high concentration antibody formulation as early as possible in the discovery/development process is desired.

Computational tools have been applied to identify drug-like antibodies that have favorable stability.[4] For viscosity prediction, Sharma et al. found that viscosity is highly correlated with variable fragment (Fv) net charge and charge symmetry and weakly correlated with hydrophobicity.[5] Based on these three parameters, a linear equation was proposed to calculate viscosity at 180 mg/ml (pH 5.5 and 200 mM arginine-HCl).[5] Spatial charge map (SCM) is another viscosity predictive tool calculated by molecular dynamics (MD) simulation that accounts for the exposed surface-negative charge distribution on the Fv region.[6] Tomer et al. proposed an equation to predict the concentration-dependent viscosity curves using charges on the heavy and light chain variable regions and the hinge region and the hydrophobic surface area of full-length antibody.[7] The comparison of these viscosity prediction tools is summarized in a recent review paper.[8] Recently, a machine learning model based on 27 mAbs was proposed to predict antibody viscosity at 150 mg/ml.[9] This machine learning model implements the decision tree (DT) classification method that includes two features of mAbs, net charge and high viscosity index (HVI). In addition, a coarse-grained model combined with hydrodynamic calculations and HVI-derived parameters were developed to predict viscosity at different concentrations.[10] For aggregation, there are several in silico models for predicting solubility/protein aggregation rates, such as Camsol,[11] Solubis,[12] and developability index (DI),[13] or identifying aggregation-prone regions, such as ANuPP,[14] Aggrescan 3D,[15] and spatial aggregation propensity (SAP).[16] The aggregation rate tools predict the kinetic rate of proteins. The aggregation-prone regions identify specific sequences that induce

CONTACT Bernhardt L. Trout trout@mit.edu Department of Chemical Engineering,Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Supplemental data for this article can be accessed on the publisher's website

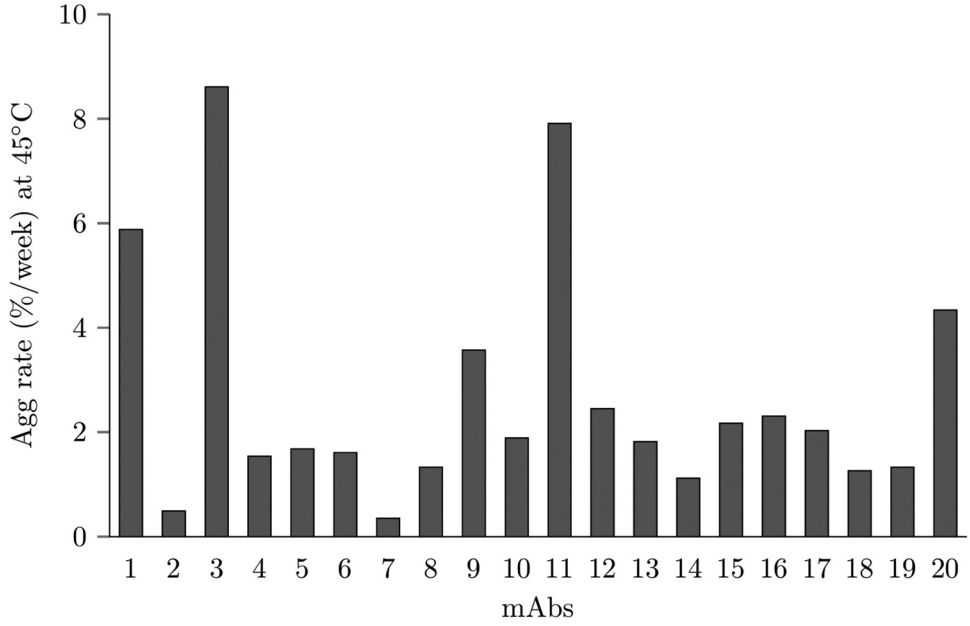

Figure 1. Aggregation rates of all 20 mAbs studied in this work.

aggregation, which can guide protein engineering to reduce the aggregation. Furthermore, machine learning has been applied to predict antibody amyloidogenesis (classification)[17–20] and protein aggregation kinetics (regression)[21,22] based on the sequence features. Antibody amyloidogenesis is of great concern for diseases in humans, but has limited application in the development of therapeutic proteins.[23] Moreover, a machine learning-based model that was trained on 21 mAbs was developed to predict therapeutic antibody aggregation rates at 150 mg/ml using structural-based features extracted from MD simulations.[24]

The molecular origin of antibody aggregation and viscosity remains unclear, but hydrophobicity and charge are considered to be the two major driving forces.[23,25] Recent studies that evaluated the aggregation and viscosity of 21 mAbs showed no overlap between those with high aggregation rates and those with high viscosity,[9,24] indicating that the underlying mechanisms of aggregation and viscosity are different. Machine learning provides a great tool to find the most relevant features for aggregation and viscosity, respectively. Previous machine learning research on predicting antibody aggregation rates and viscosity used data derived from commercial mAbs, which may have gone through molecule and/or formulation optimization for stability.[9,24] Although predictive models could be applied to mAbs in early development, such studies have not been previously reported. In this work, we measured the aggregation rates and viscosity at 150 mg/ml of 20 preclinical and clinical stage mAbs. The molecules used for this study were from a subset of preclinical/clinical stage assets that were accessible from a material generation and technology development program, with the intellectual property approved for publication purposes. Machine learning regression methods such as linear regression, support vector regression (SVR) and k-nearest neighbors (KNN) regression were applied to predict antibody aggregation rates using features obtained from MD simulations of the full-length antibody. Moreover, machine learning classification methods such as logistic regression (LR), support vector machine (SVM), KNN classification, and DT classification were implemented to predict low and high viscosity with a threshold value of 30 cP. In addition to the 20 preclinical and clinical stage mAbs in this work, we included 27 commercial mAbs from our previous work to expand the training and testing dataset. From this work, we provide here the best machine learning models as aggregation and viscosity predictive tools for antibody development.

Results

Accelerated aggregation rates

An accelerated stability study at 45°C was performed to measure aggregation of 20 mAbs in a 20 mM histidine-HCl buffer, pH 6.0 at 150 mg/mL for 2 weeks. The onset temperature (Tonset) of the first thermal transition melting temperature (Tm1) for the 20 mAbs were experimentally measured as > 50°C by differential scanning calorimetry (Table S1). Therefore, this thermal stress condition should enable an accelerated screening approach to screen the propensity for mAb aggregation, without directly imparting conformational unfolding due to storage temperature. The rate of aggregation per week is reported in Figure 1 and Table S2. Five mAbs had aggregation rates over 3% per week (mAb1, mAb3, mAb9, mAb11, and mAb20).

Machine learning and feature selection for preclinical and clinical stage antibody aggregation rates.

We applied a machine learning protocol developed from our previous work[24] to predict antibody aggregation. Thirty-five structural descriptors, including solvent-accessible surface area of hydrophobic residues (SASA_phobic), solvent-accessible surface area of hydrophilic residues (SASA_philic), SAP, spatial negative charge map (SCM_neg) and spatial positive charge map (SCM_pos) on the complementarity-

Table 1. List of mAb properties and domains for feature selection of antibody aggregation rate. The CDR definitions are based on Chothia numbering. The feature properties are obtained from dynamic average of MD trajectories. In total, there are 35 features for selection.

Feature list (mAb properties (5) x domains (7) = 35)			
mAb properties	description	domains	description
Solvent accessible surface area of hydrophobic residues (SASA_phobic)	Calculated by VMD	CDRH1	H26-H32
Solvent accessible surface area of hydrophilic residues (SASA_philic)	Calculated by VMD	CDRH2	H52-H56
Spatial aggregation propensity (SAP)	In-house program	CDRH3	H95-H102
Spatial negative charge map (SCM_neg)	In-house program	CDRL1	L24-L34
Spatial positive charge map (SCM_pos)	In-house program	CDRL2	L50-L56
		CDRL3	L89-L97
		Fv	H1-H113 + L1-L107

determining region (CDR) loops and Fv region, were used for feature selection and model building (Table 1). Surface-exposed hydrophobicity, charge patches and area have been found to correlate with antibody aggregation,[13,16,24] although the detailed mechanisms remain unknown. These could have compound effects for aggregation; therefore, all the relevant features were included for selection.

After the preprocessing step, four features were removed because of a high correlation ($r > 0.8$) with other features shown in the Supporting Information (aggregation_feature_correlation_SI.xlsx). These are SAP_pos_L2, SAP_pos_L3, SASA_phobic_H1, and SASA_phobic_L1. Exhaustive one-feature and two-feature combinations using different regression models were performed to select high-performance features based on mean square error (MSE). The MSE are averaged from 100 randomly generated fourfold cross-validation sets. Table 2 lists the top 5 one-feature and two-feature combinations using linear regression, SVR and KNN models; the complete list is in the Supporting Information (aggregation_exhaustive_SI.xlsx). For the linear model, the best one-feature is SCM_neg_H2 (MSE = 5.04), and the best two-feature combination is SCM_neg_H2 and SASA_phobic_H3 (MSE = 4.81). For the SVR model, the best one-feature is SCM_pos_H2 (MSE = 4.96), and the best two-feature combination is SCM_pos_H2 and SASA_phobic_Fv (MSE = 4.12). For the KNN model, the best one-feature and two-feature combinations are the same as that for the SVR model; however, the MSE, which are 4.35 and 3.37, respectively, are much better. Overall, the KNN model is the best for predicting aggregation rates.

Table 2. Mean squared error (MSE) of the top five one-feature and two-feature combinations of the linear regression, support vector regression (SVR) and k-nearest neighbors regression (KNN) models for predicting aggregation rates. There are 20 mAbs in this study. The MSE are averaged from 100 randomly generated fourfold cross-validation sets.

	One-feature	MSE	Two-features		MSE
Linear	SCM_neg_H2	5.04	SCM_neg_H2	SASA_phobic_H3	4.81
	SAP_pos_H1	5.31	SCM_neg_H2	SASA_philic_L3	4.97
	SASA_phobic_H3	5.49	SAP_pos_L1	SCM_neg_H2	5.08
	SCM_neg_H1	5.66	SCM_neg_H1	SASA_phobic_H3	5.19
	SASA_philic_L3	5.70	SCM_neg_H2	SCM_pos_L1	5.23
SVR	SCM_pos_H2	4.96	SCM_pos_H2	SASA_phobic_Fv	4.12
	SCM_neg_H2	5.14	SAP_pos_L1	SCM_pos_H2	4.68
	SCM_pos_L3	5.43	SAP_pos_L1	SCM_neg_H2	4.89
	SASA_phobic_Fv	5.44	SAP_pos_Fv	SASA_phobic_Fv	4.90
	SAP_pos_L1	5.46	SCM_pos_H2	SCM_pos_L3	4.90
KNN	SCM_pos_H2	4.35	SCM_pos_H2	SASA_phobic_Fv	3.37
	SCM_pos_L3	4.97	SAP_pos_L1	SCM_pos_H2	3.80
	SCM_neg_H1	5.35	SCM_neg_H1	SCM_pos_H2	3.97
	SCM_pos_H1	5.59	SCM_pos_H2	SASA_philic_L3	4.21
	SAP_pos_Fv	5.65	SCM_pos_L3	SASA_philic_L1	4.73

Cross-validation for aggregation rate models

The performance of different regression models is evaluated by the leave-out-one-cross-validation (LOOCV) method. Figure 2 illustrates the linear correlation coefficients of the experimental aggregation rates and predicted rates using the best two-feature combination from the three regression models. If the correlation coefficients of LOOCV is similar to that using the whole dataset, it indicates the predictive models can be applied in predictive models for new datasets. The correlation coefficients and root mean square errors (RMSE) of the linear regression model using all 20 data and LOOCV are 0.54 and 1.88 (%/week) and 0.38 and 2.15 (%/week), respectively. The correlation coefficients and RMSE of the SVR model using all 20 data and LOOCV are 0.88 and 1.71 (%/week) and 0.69 and 1.99 (%/week), respectively. The correlation coefficients of the KNN model using all 20 data and LOOCV are 0.89 and 1.07 (%/week) and 0.79 and 1.50 (%/week), respectively. In addition, Table 3 shows the bootstrapping results of the best two-feature combinations for the three regression models. The values of correlation coefficients (0.54, 0.88, 0.89) and RMSE (1.88, 1.71, 1.07) of the regression equations using all 20 data fall within the range of standard deviation obtained from the bootstrap method ($r = 0.56 \pm 0.12$, 0.87 ± 0.07, 0.90 ± 0.07 and RMSE 1.72 ± 0.42, 1.52 ± 0.29, 0.89 ± 0.22) for the linear, SVR and KNN models, respectively. Overall, the KNN model gives the best result for predicting antibody aggregation rates from the validation testing.

Predictive models for aggregation rates

The best model for predicting aggregation rates for preclinical and clinical antibodies is the KNN model with two features, SCM_pos_H2 and SASA_phobic_Fv. Unlike linear models whose parameters are constants, the parameters for the KNN models depend on the values of the training and testing data. It is nontrivial to show the KNN models in a concise form. Therefore, the input data for the 20 antibodies are provided in the Supporting Information (aggregation_features_SI.csv) for constructing the models, which can be used to predict the aggregation rates of new antibodies. Note that we limited the models to two features so as not to overfit.

Viscosity and diffusion interaction coefficients measurements

The viscosity measurements were conducted from 80 to 250 mg/mL at multiple shear rates depending on the mAbs tested. For mAbs that exhibit shear thinning effect, the

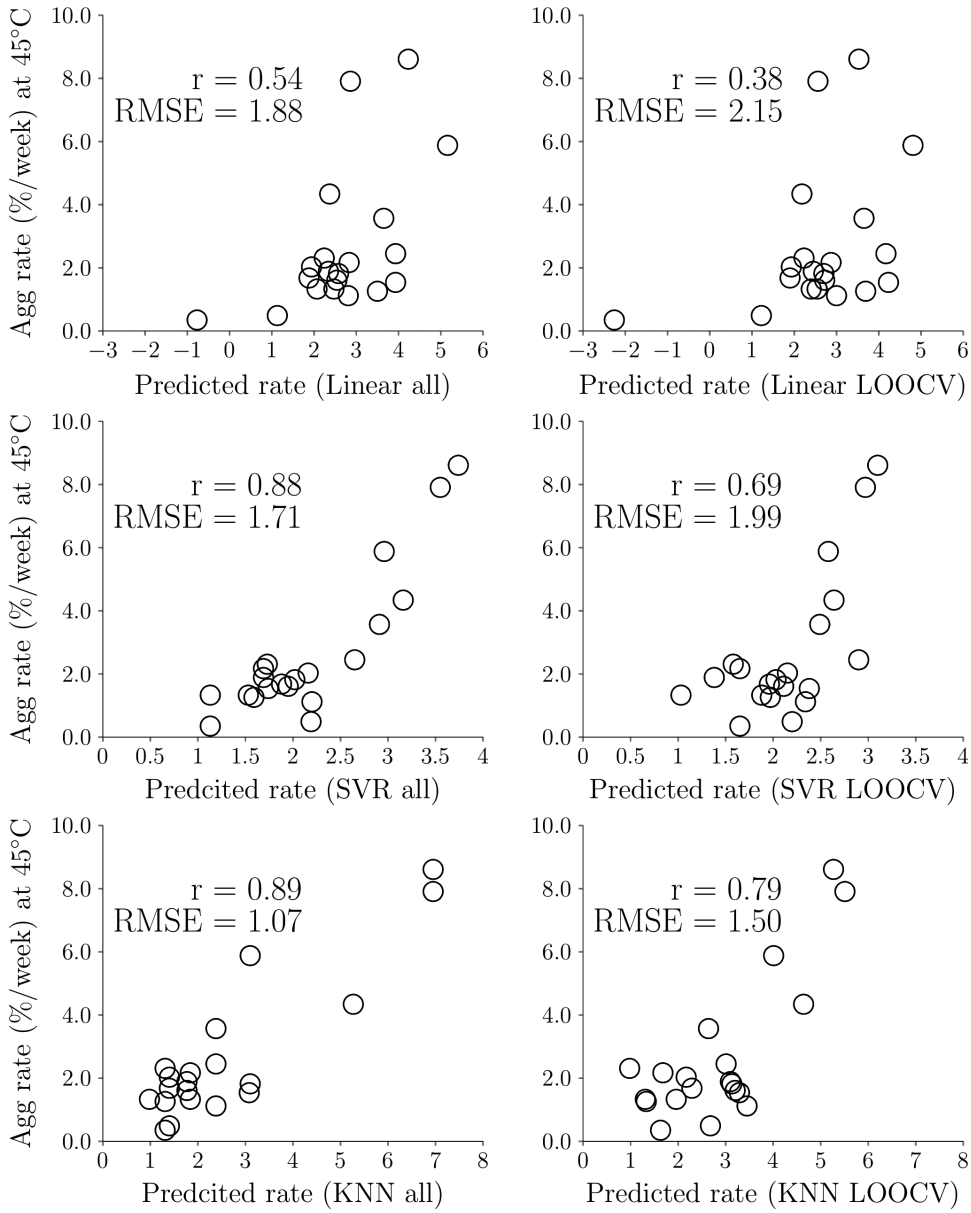

Figure 2. Correlation coefficients for the best two-feature linear, support vector regression (SVR) and k-nearest neighbors (KNN) regression models trained using all 20 data and LOOCV. The features for the linear regression model are SCM_neg_H2 and SASA_phobic_H3. The features for the SVR and KNN models are both SCM_pos_H2 and SASA_phobic_Fv.

Table 3. Bootstrapping of the best two-feature combinations for the Linear, SVR and KNN regression models. In bootstrapping, the 20 data from the original dataset were randomly sampled with replacement. The regression models were generated 100 times and average value of the regression coefficients (r), RMSE and their standard deviations were calculated.

	Two-features		r	RMSE
Linear	SCM_neg_H2	SASA_phobic_H3	0.56 ± 0.12	1.72 ± 0.42
SVR	SCM_pos_H2	SASA_phobic_Fv	0.87 ± 0.07	1.52 ± 0.29
KNN	SCM_pos_H2	SASA_phobic_Fv	0.90 ± 0.07	0.89 ± 0.22

viscosity is extrapolated to zero-shear rate at different concentrations. Figure 3 depicts the viscosity interpolated at 150 mg/ml for the 20 mAbs in this study. Six mAbs exhibit high viscosity (> 30 cP), including mAb10, mAb12, mAb13, mAb14, mAb16 and mAb20. In addition, Figure 4 plots the relationship between viscosity and diffusion interaction coefficients (kD). Five high viscosity mAbs have kD values < −5 mL/g (mAb10, mAb12, mAb13, mAb16, and mAb20). Interestingly, mAb8, which has the most negative kD value (−36 mL/g), only exhibits moderate viscosity (16.07 cP) at 150 mg/mL.

Previous viscosity predictive models

SCM scores and a decision tree model have been applied to predict or classify antibody viscosity.[6,9] These two approaches are used to predict viscosity of the 20 mAbs in this study, as shown in Table 4. The high viscosity for the predicted models is defined as SCM_neg_Fv > 1000, 12< mAb_chg<32 and HVI > 17.3 for the SCM model[6] and the machine learning model,[8] respectively. Assuming the high and low viscosity are positive and negative cases, respectively, the accuracy for the SCM model is 0.60 and the

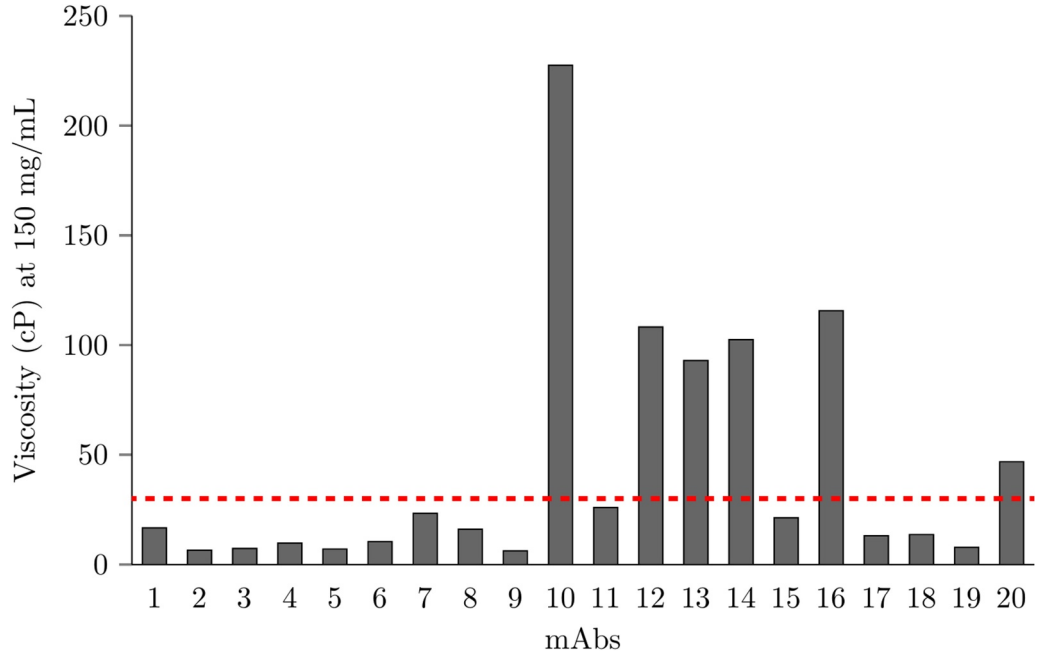

Figure 3. Viscosity at 150 mg/mL at pH 6.0 in histidine buffer of all 20 mAbs studied in this work. The red dashed line indicates the low/high viscosity cutoff (30 cP). A histogram showing the experimental viscosity at 150 mg/ml of 20 mAbs. The viscosity of mAb10, mAb12, mAb13, mAb14, mAb16 and mAb20 are above the high viscosity threshold 30 cP.

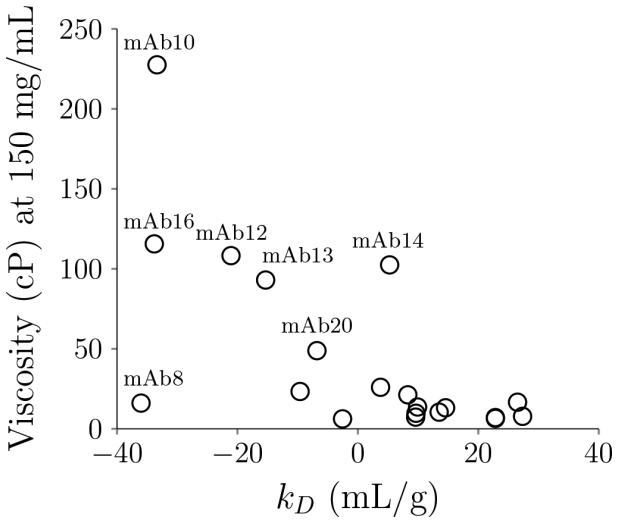

Figure 4. The relationship of viscosity at 150 mg/ml with the diffusion interaction coefficients (kD) for the 20 mAbs in this study. Open circles showing the viscosity on the y-axis and kD on the x-axis. Five high viscosity mAbs have kD values < −5 mL/g (mAb10, mAb12, mAb13, mAb16 and mAb20).

precision, recall and F1-score for the SCM model are 0.38, 0.50 and 0.43, respectively. The accuracy for the decision tree model is 0.55. The precision, recall and F1-score for the decision tree model are 0.20, 0.17 and 0.18, respectively. It should be noted that the mAb_chg criterion was 12< mAb_chg<34 in the previous work.[9] In this study, the upper-bound charge is modified to 32 because there were no low viscosity data that have mAb_chg equal to 32 in the previous study. The new criterion does not affect the performance of the previous dataset, but can correctly predict mAb4 and mAb17 as low viscosity mAbs in this work. Based on these metrics, the SCM model predicts better than that of the decision tree model for the 20 datasets. The performance of the decision tree model for the preclinical and clinical-stage antibody data is worse than that of the commercially available antibody data reported elsewhere.[2] Because the decision tree model was trained using the commercial antibody, which have different molecular origins compared to the clinical-stage antibody. Therefore, it does not generalize well to the clinical-stage antibody data.

Machine learning and feature selection for preclinical and clinical stage antibody viscosity

Commercially available antibodies are likely to have gone through stability optimization processes prior to lead candidate molecule selection. Some unstable molecular regions may have been removed. Using data on these marketed mAbs for model training is not ideal for predicting preclinical and clinical-stage antibody viscosity. In this study, the machine learning protocol we previously proposed was applied to develop new predictive models for the preclinical and clinical stage mAb data. These molecules include 18 IgG1 and 2 IgG4P isotypes, with an even distribution of lambda and kappa light chains (Table S1). Five of them have high viscosity (>30 cP) at 150 mg/ml. The commercial mAb dataset include 21 IgG1, 4 IgG2 and 2 IgG4 isotype mAbs, with 1 lambda and 26 kappa light-chain molecules.[9] Six of them have high viscosity at 150 mg/ml. The decision tree model trained from the imbalanced number of kappa and lambda light chains for the commercial mAbs could be the reason for low accuracy when predicting the preclinical/clinical stage molecules, which contain 10 IgG1 mAbs with lambda light chain (Table 4). Additionally, comparing the molecular descriptors of commercial and early-stage mAbs, we found SAP_pos_Fv is statistically different between these two groups (Table S3), which may have gone through

Table 4. Viscosity classification accuracy (ACC) of the 20 mAbs in this study using the SCM score and the machine learning model from a previous work. Predicted and experimental high viscosity are shaded in gray. The high viscosity is defined as SCM_neg_Fv > 1000, 12< mAb_chg<32 and HVI>17.3, and Vis_exp > 30 cP, respectively. Correct predictions are labeled as 1, and wrong predictions are labeled as 0.

	SCM_neg_Fv	mAb_chg	HVI	Vis_exp (150 mg/ml)		SCM_pred	ML_pred
mAb1	772.7	28	14.60	16.64		1	1
mAb2	1214.6	20	15.56	6.49		0	1
mAb3	869	28	17.02	7.30		1	1
mAb4	870	32	23.38	9.72		1	1
mAb5	1055	24	19.74	7.03		0	0
mAb6	507.6	28	14.98	10.41		1	1
mAb7	1010.2	26	16.09	23.33		0	1
mAb8	2156.2	4	12.45	16.07		0	1
mAb9	808.1	12	10.04	6.23		1	1
mAb10	667.5	24	16.74	227.54		0	0
mAb11	987.2	24	18.26	25.95		1	0
mAb12	767	30	13.79	108.25		0	0
mAb13	1089.1	22	20.18	93.00		1	1
mAb14	993.9	24	17.11	102.46		0	0
mAb15	993.6	26	20.6	21.26		1	0
mAb16	1151.8	18	16.45	115.60		1	0
mAb17	763	32	22.52	13.14		1	1
mAb18	886.9	24	12.72	13.63		1	1
mAb19	1294.7	26	22.47	7.80		0	0
mAb20	1292.7	20	16.59	48.86		1	0
					ACC (%)	60	55

different screening and optimization procedures. These are the rationales for developing new predictive models for the preclinical/clinical stage mAbs.

Table 5 lists the 35 features used for selection and model construction. These features include the number of hydrophobic residues (N_phobic), the number of hydrophilic residues (N_philic), the number of positive residues (N_pos), the number of negative residues (N_neg), net charges, charge symmetric parameter (CSP), SAP, SCM_neg, SCM_pos, and HVI on all or some of the heavy chain variable region (VH), light chain variable region (VL), Fv and mAb domains. Four different classification algorithms (LR, SVM, KNN and DT) were used to select features and evaluate model performance using exhaustive one-feature and two-feature combinations. The model performance is evaluated by accuracy (ACC) and the area under precision-recall curve (AUPRC). The ACC and AUPRC are averaged from 100 randomly generated 4-fold cross-validation sets.

Table 6 summarizes the classification results for different models. The complete list is in the Supporting Information (viscosity_exhaustive_SI.xlsx). The ACC and AUPRC for the baseline model are 0.70 and 0.30, respectively. The ACC and AUPRC of the best one-feature combinations for the four classification models range from 0.73 to 0.79 and from 0.47 to 0.59, respectively, showing slight improvement compared to that of the baseline model. However, the ACC and AUPRC of the best two-feature combinations for the four models range from 0.83 to 0.86 and from 0.64 to 0.74, respectively, which are significantly better than that of the baseline model.

Table 5. List of mAb properties and domains for feature selection of antibody viscosity. The structural features (SAP, SCM pos and SCM neg) are obtained from dynamic average of MD trajectories. Other features are extracted from antibody sequences. Charge symmetry parameters are calculated for Fv and mAb domains (2). High viscosity index is calculated for Fv domain (1). The remaining properties are calculated for VH, VL, Fv and mAb domains (8x4 = 32). In total, there are 35 features for selection.

Feature list			
mAb properties	description	domains	description
Number of hydrophobic residues (N_phobic)	A,F,I,L,M,P,V,W	VH	H1-H113
Number of hydrophilic residues (N_philic)	S,T,N,Q,Y,K,R,H,D,E	VL	L1-L107
Number of positive residues (N_pos)	K,R,H	Fv	H1-H113 + L1-L107
Number of negative residues (N_neg)	D,E	mAb	Full length
Net charges	Calculated by PROPKA3		
Charge symmetric parameter (CSP)	Product of heavy and light chain charge		
Spatial aggregation propensity (SAP)	In-house program		
Spatial positive charge map (SCM_pos)	In-house program		
Spatial negative charge map (SCM_neg)	In-house program		
High viscosity index (HVI)	In-house program		

Predictive models for preclinical and clinical stage antibody viscosity

The predictive models for the LR and DT models based on the 20 preclinical and clinical-stage antibodies are provided to classify low/high viscosity for new data. The high viscosity threshold is above 30 cP. The LR model is

$$\text{High viscosity} : -1.1 * \text{SCM\_neg\_VH} + 1.3*\text{SCM\_neg\_VL} - 0.86 > 0$$

The features need to be scaled by their means and standard deviations. The mean and standard deviation for SCM_neg_VH are 540.81 and 241.47, respectively. The mean and standard deviation for SCM_neg_VL are 466.72 and 155.21, respectively. If the predictive model is greater than 0, it is predicted to be high viscosity. Moreover, the DT model is

$$\begin{cases} Low viscosity : SAP\_pos\_VL \leq 38.05 \\ \begin{cases} Low viscosity : SAP\_pos\_VL 38.05 and N\_phobic\_VL 39.50 \\ High viscosity : SAP\_pos\_VL 38.05 and N\_phobic\_VL \leq 39.50 \end{cases} \end{cases}$$

Table 6. Accuracy (ACC) and area under the precision-recall curve (AUPRC) of the top five one-feature and two-feature combinations of the logistic regression (LR), support vector machine (SVM), k-nearest neighbors (KNN) and decision tree (DT) models for classifying low/high viscosity. There are 20 mAbs in this study. The ACC and AUPRC are averaged from 100 randomly generated 4-fold cross-validation sets. The baseline ACC is 0.70 and the baseline AUPRC is 0.30.

	One-feature	ACC	AUPRC	Two-features		ACC	AUPRC
	N_neg_VH	0.79	0.57	SCM_neg_VH	SCM_neg_VL	0.86	0.70
	SCM_neg_VL	0.77	0.54	N_neg_VH	SCM_neg_VL	0.84	0.68
LR	net charges_VH	0.78	0.53	N_neg_VH	net charges_VL	0.83	0.67
	N_neg_VL	0.77	0.51	SCM_neg_VL	SCM_pos_VH	0.83	0.66
	net charges_VL	0.74	0.48	net charges_VH	net charges_VL	0.81	0.65
	N_neg_VH	0.76	0.47	N_philic_VH	SAP_pos_VL	0.82	0.64
	net charges_VH	0.74	0.46	N_philic_Fv	SAP_pos_VL	0.82	0.63
SVM	SCM_neg_VL	0.72	0.45	N_philic_Fv	N_neg_VH	0.82	0.60
	mAbCSP	0.74	0.37	N_phobic_VL	N_neg_VH	0.82	0.60
	N_neg_VL	0.70	0.34	N_philic_VH	N_neg_VH	0.81	0.58
	HVI	0.76	0.59	N_pos_VL	N_neg_VH	0.83	0.66
	SAP_pos_VL	0.82	0.65	N_philic_Fv	FvCSP	0.82	0.64
KNN	SCM_neg_VL	0.74	0.52	N_pos_VL	net charges_VH	0.83	0.64
	net charges_VH	0.78	0.51	SCM_neg_VH	SCM_neg_VL	0.82	0.62
	N_neg_VH	0.78	0.5	N_philic_VH	FvCSP	0.76	0.62
	SAP_pos_VL	0.73	0.52	N_phobic_VL	SAP_pos_VL	0.84	0.74
	net charges_VH	0.77	0.51	N_neg_Fv	SCM_pos_VL	0.78	0.60
DT	N_neg_mAb	0.79	0.51	N_neg_mAb	net charges_VL	0.77	0.60
	SCM_pos_VL	0.75	0.49	N_neg_mAb	SCM_neg_VL	0.76	0.58
	N_neg_VH	0.76	0.49	N_phobic_VL	net charges_VH	0.79	0.56

The feature values are not scaled. The predictive models for SVM and KNN can be obtained by training the 20 mAb data using the corresponding best two-feature combinations in the Supporting Information (viscosity_features_SI.csv).

Machine learning and feature selection for combined preclinical and clinical stage and commercial antibody viscosity

One of the major challenges in applying machine learning to predict antibody stability at high concentration is developing robust models with a limited amount of data. In previous work, our group has trained a viscosity classification model using 27 commercial mAbs.[9] In this study, the viscosity of 20 preclinical and clinical stage mAbs were measured in a similar solution condition as that of the previous work (histidine/histidine-HCl buffer at pH 6.0, without surfactant and other excipients). A Chinese hamster ovary expression system used to produce the material and, following purification, the starting monomer purity was >95%. In both studies, the viscosity was measured at 18–20°C by VROC Initium viscometer at multiple shear rates. For the 27 commercial mAbs, non-Newtonian effects were assumed. In this study, we found non-Newtonian effects for low viscosity mAbs were negligible, but significant for high viscosity mAbs. For high viscosity mAbs, viscosity was extrapolated to zero-shear rate.

In order to expand the data size, we combined the two datasets for machine learning training. In total, there are 47 data that cover preclinical, clinical and commercial mAbs. The same protocol and features were used as those for the 20 preclinical and clinical mAbs described previously. Table 7 shows the top 5 one-feature and two-feature combinations for different classification models. The complete list is in the Supporting Information (viscosity_exhaustive_combined_SI.xlsx). The ACC and AUPRC for the baseline model are 0.74 and 0.26, respectively. The best one-feature for the LR, SVM and DT models are the same, mAbCSP, which have the same

Table 7. Accuracy (ACC) and area under the precision-recall curve (AUPRC) of the top five one-feature and two-feature combinations of the logistic regression (LR), support vector machine (SVM), k-nearest neighbors and decision tree (DT) models for classifying low/high viscosity. There are 20 mAbs in this study plus 27 mAbs from the literature. The ACC and AUPRC are averaged from 100 randomly generated 4-fold cross-validation sets. The baseline ACC is 0.74 and the baseline AUPRC is 0.26.

	One-feature	ACC	AUPRC	Two-features		ACC	AUPRC
	mAbCSP	0.81	0.49	N_phobic_VL	net charges_VL	0.85	0.60
	net charges_VL	0.76	0.39	N_phobic_VL	mAbCSP	0.85	0.58
LR	N_neg_VL	0.77	0.37	net charges_VL	HVI	0.84	0.56
	FvCSP	0.76	0.36	N_phobic_Fv	net charges_VL	0.84	0.56
	N_pos_VL	0.75	0.35	N_phobic_mAb	net charges_VL	0.83	0.55
	mAbCSP	0.81	0.47	N_phobic_VL	net charges_VL	0.83	0.53
	net charges_mAb	0.77	0.37	N_philic_mAb	mAbCSP	0.83	0.51
SVM	net charges_VL	0.76	0.37	net charges_mAb	mAbCSP	0.83	0.50
	N_pos_VL	0.73	0.29	N_neg_VH	net charges_mAb	0.82	0.49
	net charges_VH	0.75	0.28	net charges_VL	net charges_mAb	0.82	0.49
	net charges_mAb	0.78	0.47	N_neg_Fv	net charges_VL	0.85	0.57
	N_phobic_VH	0.77	0.42	net charges_VL	net charges_mAb	0.82	0.53
KNN	net charges_VL	0.78	0.42	net charges_VH	net charges_mAb	0.82	0.53
	mAbCSP	0.76	0.41	N_philic_VL	net charges_VL	0.82	0.53
	SAP_pos_VL	0.73	0.39	mAbCSP	HVI	0.80	0.53
	mAbCSP	0.81	0.47	N_phobic_VL	net charges_VL	0.85	0.57
	SAP_pos_mAb	0.75	0.41	net charges_VL	net charges_mAb	0.84	0.56
DT	net charges_mAb	0.75	0.40	N_philic_VL	net charges_VL	0.84	0.54
	net charges_VL	0.76	0.39	SAP_pos_mAb	FvCSP	0.78	0.48
	net charges_VH	0.76	0.35	SCM_pos_VL	mAbCSP	0.80	0.48

ACC (0.81) and similar AUPRC (0.47 to 0.49). Similarly, the best two-feature combinations for the LR, SVM and DT models are also the same, N_phobic_VL and net charges_VL (ACC = 0.83 to 0.85 and AUPRC = 0.53 to 0.60). On the other hand, the best one-feature for the KNN model is net charges_mAb (ACC = 0.78; AUPRC = 0.47). The best two-feature combination for the KNN model is N_neg_Fv and net charges_VL (ACC = 0.85; AUPRC = 0.57).

Predictive models for antibody viscosity from combined datasets

The predictive models for the LR and DT models using the 20 preclinical and clinical stage and the 27 commercial antibodies are provided to classify low/high viscosity for new data. The LR predictive model is

$$\text{High viscosity}: -0.72 * \text{N\_phobic\_VL} - 1.17 * \text{netcharges\_VL} - 1.19 > 0$$

The features need to be scaled by their means and standard deviations. The mean and standard deviation for N_phobic_VL are 37.91 and 2.70, respectively. The mean and standard deviation for net charges_VL are 0.64 and 1.93, respectively.

The DT predictive model is

$$\begin{gathered} Low viscosity: netcharges\_VL > -0.50 \\ Low viscosity: netcharges\_VL \leq -0.50 and N\_phobic\_VL > 38.0 \\ High viscosity: netcharges\_VL \leq -0.50 and N\_phobic\_VL \leq 38.0 \end{gathered}$$

Similarly, these feature values are unscaled. In addition, the predictive models for SVM and KNN can be constructed by training the 47 mAb data using the best two-feature combination in the Supporting Information (viscosity_features_combined_SI.csv).

Discussion

In this study, we measured the aggregation rates and viscosity of 20 preclinical and clinical stage mAbs at high concentration. Antibodies having the top 5 highest aggregation rates are mAb1, mAb3, mAb9 and mAb11 and mAb20, and the top 5 highest viscosity are mAb10, mAb12, mAb13, mAb14 and mAb16. Interestingly, these groups of mAbs do not overlap, suggesting that the driving forces for antibody aggregation and viscosity may be different.

Antibody self-association is considered to promote high viscosity.[26] Diffusion interaction coefficients (kD) are commonly used to measure protein–protein interactions, although their relationship to predict viscosity remains controversial.[27,28] Figure 4 shows that most high viscosity mAbs have large negative kD values; however, mAb8, which has the most negative kD value exhibits low viscosity. From the SCM score in Table 4, mAb8 has the highest SCM score, indicating strong electrostatic interactions due to negative charge patches on the Fv region, which supports the experimental kD measurement. Kingsbury et al. recently found that antibody solutions that have large negative kD values could exhibit either high viscosity or high opalescence.[27] We found that mAb8 also exhibits high solution opalescence, which agrees with the previous finding. Although kD or the SCM score cannot distinguish high viscosity and high opalescence, they are still good indicators for poor stability.

The protocol and machine learning features described in this paper are built on our previous works.[6,9,16,24] Because of the limited availability of high concentration therapeutic antibody aggregation and viscosity data, it is of great value to evaluate the performance of existing models and improve the predictive models using larger datasets.

We applied machine learning to predict the antibody aggregation rates at 45°C in a 20 mM histidine-HCl buffer, pH 6.0 at 150 mg/mL based on 20 preclinical and clinical stage mAbs. It is worth noting that the ranking of aggregation tendency may differ and the accelerated thermal stress conditions may not always correlate to real-time stability at the intended storage conditions. This may be due to differences in the molecular origins of degradation pathways, impacting the physicochemical stability and resulting in conformational changes of the protein structure.[29–31] The accelerated stability condition described here provides a screening approach to assess the propensity for aggregation, especially in a controlled matrix (i.e., base buffer, with no stabilizing excipients). Of course, the approach that we present here can be used to parameterize the model at any conditions. In previous work, we developed an aggregation rate model at 40°C in a 10 mM histidine-HCl buffer, pH 6.0 at 150 mg/mL based on 21 commercial antibodies.[24] Although the solution conditions (buffer and pH) are similar for the two datasets, the difference in the temperature makes the aggregation rates very different even for the same antibody (data not shown). As such, the previous model may not be directly applicable to the data at 45°C. Therefore, new models were built using a similar protocol as the previous work.[24] In this study, we found the best aggregation rates model is the KNN model, which agrees with our previous work.[24] The best two-feature combination of the KNN model is SCM_pos_H2 and SASA_phobic_Fv. Hydrophobicity has been used to predict antibody aggregation in earlier works.[13,16] In addition, in our previous work, we also found SCM_pos is an important feature for the antibody aggregation rate. It should be noted that SCM includes a distance cutoff of 10 Å, so SCM_pos_H2 does not mean only the positive charges on the CDRH2 region are important. Residues surrounding CDRH2 should be also considered.

It has been suggested that, due to their complex nature, mAb degradation pathways may or may not follow Arrhenius behavior/kinetics. Therefore, expanding this approach to extrapolate to real-time storage for predicting shelf-life considerations could be difficult. Recently, Kuzman et al. and Gentiluomo et al. have shown some potential for predicting long-term shelf-life stability when using non-linear machine learning models.[32,33] This, however, also leads to some challenges when assuming first-order kinetics of proteins that may be susceptible to different degradation pathways that may affect chemical and physical stability upon exposure to thermal

stress.[34] Therefore, the accelerated stability and real-time stability measurements are different approaches. Of course, long-term data could easily be used to train our model using our methodology.

Two sets of antibody viscosity data for 47 mAbs in total were used for training and testing. In this study, the viscosity of 20 preclinical and clinical stage mAbs were measured. In our previous study, the viscosities of 27 commercial antibodies were measured. We performed a blind test by using the ML model, trained from the 27 commercial antibodies, to predict the 20 investigational antibodies in this study (Table 4), and the results were not satisfactory. Because of the limited dataset, training on a subset of data is prone to find features not the most relevant for viscosity. In order to generalize our predictive models, we decided to combine both datasets for the ML algorithms to capture the common features. The experimental conditions are very similar (pH = 6.0 in 10–20 mM histidine-HCl buffer at 18–20°C), and for binary classification, slight viscosity variation from equipment setup and operation do not change the overall low/high viscosity categories. The DT model obtained from the commercial mAbs were applied to predict preclinical and clinical stage mAb data. The accuracy was only 0.55, indicating that the underlying mechanism of the preclinical and clinical mAbs could be different from the marketed mAbs so that the DT model does not capture these features. By using the same protocol, new predictive models based on the 20 new data were developed. The performance for these classification models is similar (ACC = 0.82 to 0.86; AUPRC = 0.64 to 0.74), although the best two-feature combination for each model varies. Conversely, the best LR, SVR and DT models for the combined 47 datasets share the same one-feature and two-feature combinations. As the number of datasets increases, only the most important feature combinations are selected despite the statistical models implemented. The best two features are N_phobic_VL and net charges_VL. Both hydrophobicity and net charges are reported to be related to antibody viscosity. The machine learning models provide a quantitative relationship to connect them with antibody viscosity. These two features are sequence-based descriptors, which can be implemented very efficiently. Why only the VL regions matter to the viscosity prediction is still unknown. More data are needed to validate these models.

Although there are some public databases for antibody aggregation and protein aggregation kinetics such as CPAD 2.0,[35] these data focus primarily on amyloid aggregates. These amyloid aggregates are related to immunogenicity in animal models, but are of limited utility for pharmaceutical proteins.[23] Currently, there is no public database available for therapeutic antibody aggregation rates and viscosity at high concentrations. Data from published literature could be performed in different solution conditions (pH, buffers, excipients, and protein concentrations), but very often the sequence information is not available. This is one of the major challenges for applying machine learning to predict antibody stability. In addition, because high concentration antibodies are expensive to produce, it is not feasible to obtain a large number from one source with sufficient amount of data for machine learning applications. Combining the datasets from different sources with proper pre-experimental designs as performed in this study is a possible solution.

Materials and methods

Protein preparation

The 20 mAbs used for this study were internally manufactured at AstraZeneca (Gaithersburg, MD) and consisted of a combination of 18 IgG1and 2 IgG4P subclass mAbs (Table S1). The protein solutions were obtained as bulk Drug Substance, in a molecule-respective, non-surfactant containing, formulation buffer. The starting monomer purity for each mAb was >95% as measured by high-performance size exclusion chromatography (HPSEC; Agilent Technologies Santa Clara, CA) using a TSK-Gel G3000SWXL HPLC column (Tosoh Bioscience LLC, Montgomeryville, PA) and mobile phase comprised of 0.1 M sodium phosphate dibasic anhydrous, 0.1 M sodium sulfate, and 0.05 M sodium azide at pH 6.8 with 250 µg protein injection. The mAbs solutions were individually buffer exchanged into a formulation buffer of 20 mM histidine-HCl at pH 6.0 using 10 K MWCO Slide-A-Lyzer dialysis cassette (Thermo Scientific). Dialysis was performed overnight with multiple buffer exchanges at a minimum buffer-to-protein solution ratio of 1000:1. The dialyzed product was tested to meet the appropriate pH and osmolality requirements. The samples were then concentrated using Amicon Ultra-4 Centrifugal Filter units with 10 K MWCO (EMD Millipore, Merck KGaA, Darmstadt, Germany) to a target concentration of 150 mg/mL. Total protein was measured using a UV-vis spectrophotometer (Trinean DropSense 96, Unchained Labs Pleasanton, CA) with respective mAb experimentally determined extinction coefficients and corrected for density when necessary.

Measurement of accelerated aggregation rates

Samples were 0.22 µm filtered (PVDF membrane, EMD Millipore, Merck KGaA, Darmstadt, Germany) and aseptically hand filled into 2 R glass vials (Std Type 1, USP; Schott) with rubber stoppers (13 mm chlorobutyl, Diakyo/West Pharmaceutical Services) and aluminum overseals (13 mm, West Pharmaceutical Services). Samples were placed in a temperature and humidity-controlled incubation chamber with setpoints of 45°C and 75% relative humidity. The vials were aseptically sampled on 2-day intervals for a total duration of 2 weeks. The pulled samples were prepared for HPSEC analysis by diluting to 10 mg/mL with 0.2 µm filtered formulation buffer and 250 µg of protein injected (similar to method described above). The rate of aggregation was determined using linear regression of the total content of aggregates over the timecourse of the stability study (Table S2).

Measurements of viscosity

Viscosity was measured at multiple concentrations (3–6 concentrations each construct) ranging from 80 mg/mL to 250 mg/mL dependent on mAb sample; all samples included at least one measurement of concentrations >150 mg/mL aside from mAb 11, which was measured at a highest concentration of 142 mg/mL due to material constraints. All mAb samples were formulated in 20 mM histidine-HCl buffer at pH 6.0. Prior to

concentration and viscosity measurement, samples were passed through a 0.45 μm filter. Concentration was determined using the UV-vis spectrophotometry method described above.

Using a VROC Initium viscometer (Rheosense, San Ramon, CA), viscosities were determined at multiple shear rates between 300 and 50,000 s^{-1} with a B05 or E02 measuring chip where appropriate to ensure optimal pressure across the sensor array of the chip. Approximate zero shear viscosity for each sample exhibiting shear thinning was estimated through extrapolation of measured viscosities across multiple shear rates. The viscosity at 150 mg/mL was then interpolated by a best fit equation of natural log of viscosity vs concentration for each construct.

Measurements of diffusion interaction parameters

The diffusion interaction parameter (k_D) was calculated using measurements obtained from experimental diffusion coefficient as a function of total protein concentration (DynaPro Plate Reader II -Wyatt, Santa Barbara, CA). Protein samples were equilibrated to room temperature and titrations were prepared at 2, 4, 6, 8, and 10 mg/mL in formulation buffer (20 mM histidine-HCl, pH 6.0) and filtered using a 0.22 um syringe filter. Using a low volume 384-well plate (Corning, Tewksbury, MA), samples were meticulously aliquot in triplicate (35 μL each). A run method protocol was written using the Dynamics software package (Wyatt, Santa Barbara, CA; version 7.1.9.3) to analyze samples at 25°C using an 830-nm laser and the sample acquisitions were set to 5 seconds, with 10 total acquisitions collected for each sample run. Correction factors such as viscosity and refractive index were also provided prior to sample analysis. The data was exported to excel and plots created to determine the slope and y-intercept for the diffusion coefficient versus total protein concentration. The k_D was subsequently calculated by taking the ratio of the slope and the y-intercept values.

Computational modeling of mAbs

The mAb molecules were constructed following the protocol proposed by Brandt et al.[36] Briefly, the structure of antigen-binding fragment (Fab) region was superimposed on a template structure obtained from the KOL/Padlan structure.[37,38] The immunoglobulin G1 (IgG1) template was obtained from the KOL/Padlan structure. For IgG4 models, the Fc regions (PDB: 4C54) were superimposed on the KOL/Padlan IgG1 structure. The Fab structure was retrieved from either available crystal structures or homology model built from RosettaAntibody.[39–41] Disulfide bridges were carefully matched to the respective isotypes. The glycosylation pattern for each mAb was modeled according to available literature data. For mAbs without literature data on the glycosylation pattern, the G_0F glycosylation pattern was chosen.

Molecular dynamics simulations

Molecular dynamics simulations were performed using all-atom structures with explicit solvent using the TIP3P water model.[42] Simulation boxes were set up using visual MD to place a single antibody in a water box extending 12 Å beyond the protein surface.[43] Simulations were performed at 300 K and 1 atm in the NPT ensemble, using the NAMD software package and the CHARMM36m force field.[44–46] The system pH was set to 6.0 to match the experimental pH by adjusting the protonation states of histidine residues using the PROPKA3 protocol.[47] Electrostatic interactions were treated with the Particle Mesh Ewald (PME) method and van der Waals interactions were calculated using a switching distance of 10 Å and a cutoff of 12 Å.[48] The integration time step was set to 2 fs. Each mAb system was pre-equilibrated for 10 ns, followed by 50 ns production runs.

Feature selection for aggregation rates and viscosity

Based on a previous study, structural features obtained from MD simulations were extracted for building regression models for aggregation rates.[24] Table 1 lists the features used for aggregation rates in this work. We included structural features such as SASA_phobic, SASA_philic, SAP and SCM_neg and SCM_pos covering 6 CDR and 1 Fv regions for selection.[24] In total, there are 35 features (see supporting information: aggregation_features_SI.csv for details). They were calculated from averaging 50 ns MD trajectories. The 50 ns simulation is long enough to obtain converged feature values, but may not capture large conformational change of antibodies. In the pre-processing step, highly correlated features (correlation coefficient > 0.8) were filtered to keep only one of them from each pair. The features for viscosity classification contain both structural and sequence descriptors (Table 5) as described previously.[9] In total, there are 35 features (see Supporting Information: viscosity_features_SI.csv for details).

For each machine learning method described in the next section, exhaustive feature selection for one-feature and two-features were performed to search for the best feature combinations using the exhaustive feature selector tool from mlxtend library.[49] The best feature combinations for the regression models were selected based on their mean squared errors. The best feature combinations for the classification models were selected based on their AUPRC. AUPRC was chosen because the dataset contains an imbalanced number of high and low viscosity antibodies.

Machine learning methods for aggregation rates and viscosity

All the machine learning methods were implemented using the scikit-learn library.[50] The protocols follow our previous works.[9,24] Briefly, different regression models were used for aggregation rates including linear regression (*linear_model.LinearRegression()*), nearest neighbors regression (*neighbors.KNeighborsRegressor()*) and support vector regression (*svm.SVR()*). For viscosity classification, logistic regression (*linear_model.LogisticRegression()*), support vector machine (*linear_model.svm()*), nearest neighbors classification (*neighbors.KNeighborsClassifier()*) and decision tree classification (*tree.DecisionTreeClassifier()*) models were employed. The functions utilized from the scikit-learn library were specified in the parentheses. The default parameters were used for all

functions, except the number of neighbors in the KNN models is 3 and the maximum depth in the DT models is 2.

Abbreviations

CDRcomplementarity-determining region
CDRH1 the first complementarity-determining region of the heavy chain
CDRH2 the second complementarity-determining region of the heavy chain
CDRH3 the third complementarity-determining region of the heavy chain
CDRL1 the first complementarity-determining region of the light chain
CDRL2 the second complementarity-determining region of the light chain
CDRL3 the third complementarity-determining region of the light chain
CSP charge symmetric parameter
DT decision tree
Fv variable fragment
HVI high viscosity index
IgG1 immunoglobulin G1
KNNk- nearest neighbors
LOOCV Leave-out-one-cross-validation
LR logistic regression
mAbs monoclonal antibodies
MD molecular dynamics
MSE mean square error
N_negnumber of negative residues
N_philicnumber of hydrophilic residues
N_phobicnumber of hydrophobic residues
N_posnumber of positive residues
RMSE root mean square error
SAP spatial aggregation propensity
SASA solvent-accessible surface area
SCM spatial charge map
SVM support vector machine
SVR support vector regression
Tonsetonset temperate
Tm1 first thermal transition melting temperature
VHheavy chain variable region
VLlight chain variable region

Acknowledgments

We would like to thank AstraZeneca for funding this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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

References

1. Whitaker N, Xiong J, Pace SE, Kumar V, Middaugh CR, Joshi SB, Volkin DB. A formulation development approach to identify and select stable ultra-high-concentration monoclonal antibody formulations with reduced viscosities. J Pharm Sci. 2017;106(11):3230–41. doi:10.1016/j.xphs.2017.06.017.
2. 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. 2018;11(2):239–64. doi:10.1080/19420862.2018.1553476.
3. Jarasch A, Koll H, Regula JT, Bader M, Papadimitriou A, Kettenberger H. Developability assessment during the selection of novel therapeutic antibodies. J Pharm Sci. 2015;104(6):1885–98. doi:10.1002/jps.24430.
4. Makowski EK, Wu L, Gupta P, Tessier PM. Discovery-stage identification of drug-like antibodies using emerging experimental and computational methods. mAbs. 2021;13(1):1895540. doi:10.1080/19420862.2021.1895540.
5. 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. 2014;111(52):18601–06. doi:10.1073/pnas.1421779112.
6. 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. 2015;8(1):43–48. doi:10.1080/19420862.2015.1099773.
7. Tomar DS, Li L, Broulidakis MP, Luksha NG, Burns CT, Singh SK, Kumar S. In-silico prediction of concentration-dependent viscosity curves for monoclonal antibody solutions. mAbs. 2017;9(3):476–89. doi:10.1080/19420862.2017.1285479.
8. Kuroda D, Tsumoto K. Engineering stability, viscosity, and immunogenicity of antibodies by computational design. J Pharm Sci. 2020;109(5):1631–51. doi:10.1016/j.xphs.2020.01.011.
9. Lai P-K, 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):1167–75. doi:10.1021/acs.molpharmaceut.0c01073.
10. Lai P-K, Swan JW, Trout BL. Calculation of therapeutic antibody viscosity with coarse-grained models, hydrodynamic calculations and machine learning-based parameters. mAbs. 2021;13(1):e1907882. doi:10.1080/19420862.2021.1907882.
11. Sormanni P, Aprile FA, Vendruscolo M. The camsol method of rational design of protein mutants with enhanced solubility. J Mol Biol. 2015;427(2):478–90. doi:10.1016/j.jmb.2014.09.026.
12. De Baets G, Van Durme J, van der Kant R, Schymkowitz J, Rousseau F. Solubis: optimize your protein. Bioinformatics. 2015;31(15):2580–82. doi:10.1093/bioinformatics/btv162. (Oxford, England)
13. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. J Pharm Sci. 2012;101(1):102–15. doi:10.1002/jps.22758.
14. Prabakaran R, Rawat P, Kumar S, Michael Gromiha M. ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. J Mol Biol. 2021;433(11):166707. doi:10.1016/j.jmb.2020.11.006. (Computation Resources for Molecular Biology)
15. Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. Nucleic Acids Res. 2019;47(W1):W300–W307. doi:10.1093/nar/gkz321.
16. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci. 2009;106(29):11937–42. doi:10.1073/pnas.0904191106.
17. Liaw C, Tung C-W, Ho S-Y, Isalan M. Prediction and analysis of antibody amyloidogenesis from sequences. PloS One. 2013;8(1):e53235. doi:10.1371/journal.pone.0053235.
18. David MPC, Concepcion GP, Padlan EA. Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. BMC Bioinform. 2010;11(1):79. doi:10.1186/1471-2105-11-79.

19. Rawat P, Prabakaran R, Kumar S, Gromiha MM. Exploring the sequence features determining amyloidosis in human antibody light chains. Sci Rep. 2021;11(1):13785. doi:10.1038/s41598-021-93019-9.
20. Garofalo M, Piccoli L, Romeo M, Barzago MM, Ravasio S, Foglierini M, Matkovic M, Sgrignani J, De Gasparo R, Prunotto M, et al. Machine learning analyses of antibody somatic mutations predict immunoglobulin light chain toxicity. Nat Commun. 2021;12(1):3532. doi:10.1038/s41467-021-23880-9.
21. Rawat P, Prabakaran R, Kumar S, Gromiha MM. AbsoluRATE: an in-silico method to predict the aggregation kinetics of native proteins. Biochim Biophys Acta Proteins Proteomics. 2021;1869(9):140682. doi:10.1016/j.bbapap.2021.140682.
22. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Sci: Publication Protein Society. 2005;14(10):2723–34. doi:10.1110/ps.051471205.
23. Roberts CJ. Therapeutic protein aggregation: mechanisms, design, and control. Trends Biotechnol. 2014;32(7):372–80. doi:10.1016/j.tibtech.2014.05.005.
24. Lai P-K, Fernando A, Cloutier TK, Kingsbury JS, Gokarn Y, Halloran KT, Calero-Rubio C, Trout BL. Machine learning feature selection for predicting high concentration therapeutic antibody aggregation. J Pharm Sci. 2021;110(4):1583–91. doi:10.1016/j.xphs.2020.12.014.
25. 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):216–28. doi:10.1080/19420862.2015.1128606.
26. Liu J, Nguyen MDH, Andya JD, Shire SJ. Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution. J Pharm Sci. 2005;94(9):1928–40. doi:10.1002/jps.20347.
27. Kingsbury JS, Saini A, Auclair SM, Fu L, Lantz MM, Halloran KT, Calero-Rubio C, Schwenger W, Airiau CY, Zhang J, et al. A single molecular descriptor to predict solution behavior of therapeutic antibodies. Science Advances. 2020;6(32):eabb0372. doi:10.1126/sciadv.abb0372.
28. Woldeyes MA, Qi W, Razinkov VI, Furst EM, Roberts CJ. How well do low- and high-concentration protein interactions predict solution viscosities of monoclonal antibodies? J Pharm Sci. 2019;108(1):142–54. doi:10.1016/j.xphs.2018.07.007.
29. Krause ME, Sahin E. Chemical and physical instabilities in manufacturing and storage of therapeutic proteins. Curr Opin Biotechnol. 2019;60:159–67. doi:10.1016/j.copbio.2019.01.014.
30. Wang W. Protein aggregation and its inhibition in biopharmaceutics. Int J Pharm. 2005;289(1–2):1–30. doi:10.1016/j.ijpharm.2004.11.014.
31. Leblanc Y, Ramon C, Bihoreau N, Chevreux G. Charge variants characterization of a monoclonal antibody by ion exchange chromatography coupled on-line to native mass spectrometry: case study after a long-term storage at +5°C. J Chromatogr B Analyt Technol Biomed Life Sci. 2017;1048:130–39. doi:10.1016/j.jchromb.2017.02.017.
32. Kuzman D, Bunc M, Ravnik M, Reiter F, Žagar L, Bončina M. Long-term stability predictions of therapeutic monoclonal antibodies in solution using Arrhenius-based kinetics. Sci Rep. 2021;11(1):20534. doi:10.1038/s41598-021-99875-9.
33. Gentiluomo L, Roessner D, Frieß W. Application of machine learning to predict monomer retention of therapeutic proteins after long term storage. Int J Pharm. 2020;577:119039. doi:10.1016/j.ijpharm.2020.119039.
34. Brummitt RK, Nesta DP, Roberts CJ. Predicting accelerated aggregation rates for monoclonal antibody formulations, and challenges for low-temperature predictions. J Pharm Sci. 2011;100(10):4234–43. doi:10.1002/jps.22633.
35. Rawat P, Prabakaran R, Sakthivel R, Mary Thangakani A, Kumar S, Gromiha MM. CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. Amyloid. 2020;27(2):128–33. doi:10.1080/13506129.2020.1715363.
36. Brandt JP, Patapoff TW, Aragon SR. Construction, MD simulation, and hydrodynamic validation of an all-atom model of a monoclonal IgG antibody. Biophys J. 2010;99(3):905–13. doi:10.1016/j.bpj.2010.05.003.
37. Padlan EA. Anatomy of the antibody molecule. Mol Immunol. 1994;31(3):169–217. doi:10.1016/0161-5890(94)90001-9.
38. Boehm MK, Woof JM, Kerr MA, Perkins SJ. 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 modelling. J Mol Biol. 1999;286(5):1421–47. doi:10.1006/jmbi.1998.2556.
39. Sivasubramanian A, Sircar A, Chaudhury S, Gray JJ. Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. Proteins. 2009;74(2):497–514. doi:10.1002/prot.22309.
40. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ. Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. Proteins. 2014;82(8):1611–23. doi:10.1002/prot.24534.
41. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, Adolf-Bryfogle J, Biswas N, Dunbrack RL, Gray JJ. Modeling and docking of antibody structures with Rosetta. Nat Protoc. 2017;12(2):401–16. doi:10.1038/nprot.2016.180.
42. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983;79(2):926–35. doi:10.1063/1.445869.
43. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996;14(1):33–38. doi:10.1016/0263-7855(96)00018-5.
44. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26(16):1781–802. doi:10.1002/jcc.20289.
45. Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD, Pastor RW. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J Phys Chem B. 2010;114(23):7830–43. doi:10.1021/jp101759q.
46. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot Bl, Grubmüller H, MacKerell AD, de Groot BL. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017;14(1):71–73. doi:10.1038/nmeth.4067.
47. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: consistent treatment of internal and surface residues in empirical PKA predictions. J Chem Theory Comput. 2011;7(2):525–37. doi:10.1021/ct100578z.
48. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995;103(19):8577–93. doi:10.1063/1.470117.
49. 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. doi:10.21105/joss.00638.
50. 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(85):2825–30.