molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

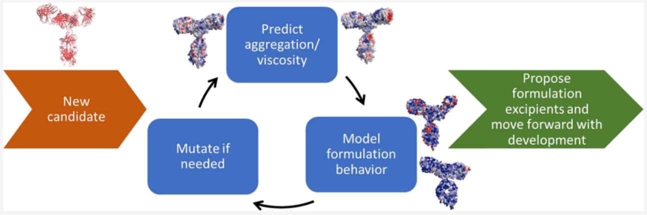
Machine Learning Models of Antibody−Excipient Preferential Interactions for Use in Computational Formulation Design

Theresa K. Cloutier, Chaitanya Sudrik, Neil Mody, Hasige A. Sathish, and Bernhardt L. Trout*

Cite This: *Mol. Pharmaceutics* 2020, 17, 3589−3599

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information


New candidate
Predict aggregation/ viscosity
Mutate if needed
Model formulation behavior
Propose formulation excipients and move forward with development

ABSTRACT: Preferential interactions of formulation excipients govern their impact on the stability properties of proteins in solution. The ability to predict these interactions without the need to perform experiments would enable formulation design to begin early in the development of a new antibody therapeutic. With that in mind, we developed a feature set to numerically describe local regions of an antibody's surface for use in machine learning applications. Then, we used these features to train machine learning models for local antibody−excipient preferential interactions for the excipients sorbitol, sucrose, trehalose, proline, arginine·HCl, and NaCl. Our models had accuracies of up to about 85%. We also used linear (elastic net) models to quantify the contribution of antibody surface features to the preferential interaction coefficients, finding that the carbohydrates and proline tend to have similar important features, while the interactions of arginine·HCl and NaCl are governed by charge features. We present several case studies demonstrating how these machine learning models could be used to predict experimental aggregation and viscosity behavior in solution. Finally, we propose an approach to computational formulation design wherein a panel of excipients may be considered while designing an antibody sequence.

KEYWORDS: *antibodies, aggregation, viscosity, preferential interaction coefficient, formulation, excipient*

■ INTRODUCTION

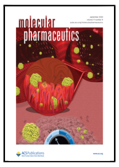
Monoclonal antibodies (mAbs) are a rapidly growing class of therapeutic proteins that are frequently formulated at high concentrations.[1,2] At high concentrations, physical instabilities, such as aggregation and viscosity, must be often addressed.[3,4] These instabilities can be mitigated by mutating the antibody so that it is more stable or by a trial and error selection of formulation excipients but often these do not work.[1] One of the reasons is that even though some general trends in the impacts of certain classes of excipients on mAb behavior have been identified, formulation design generally requires an experimental screening approach, which is costly in terms of time and material.[4]

There has been significant work in understanding how the antibody sequence and structure tend to impact aggregation and viscosity behaviors.[5−8] However, these studies rarely consider the impact of excipients possibly because the mechanisms of antibody−excipient interactions are generally not well-understood and there can be significant variations in the impact of the same excipient on different mAbs.[9−11] Timasheff and colleagues pioneered the study of protein−excipient preferential interactions, examining how preferential hydration of globular proteins affected stability.[12−15] More recently, we have measured the preferential interactions of a variety of excipients with three different mAbs.[16,17]

However, these experimental approaches are limited to studying the global or average interactions of excipients with entire proteins but do not measure the local interactions. Antibody aggregation and viscosity behaviors are thought to be driven by local antibody−antibody interactions, for example, with particular interactions of hydrophobic patches leading to reversible and eventually irreversible to aggregation events.[8] Therefore, the local interactions of excipients with this particular patch may directly affect the aggregation behavior. Even if an excipient is globally (on average) excluded from the antibody, it may be locally included near a hydrophobic surface patch and thus may have the potential to disrupt aggregation.

Although it is possible to examine certain local behaviors experimentally, such as using HD-exchange to understand how local mAb flexibility changes in the presence of an excipient,[18] molecular-level resolution of these interactions is best accessed via simulation. HD-exchange is sensitive at the peptide level,[18] while simulations provide residue-level resolution. Recently, Tilegenova et al. studied the interactions of mono- and bispecific antibodies with arginine·HCl (Arg·HCl) via experiment and simulation, identifying point

Received: June 13, 2020
Revised: August 6, 2020
Accepted: August 14, 2020
Published: August 14, 2020



https://dx.doi.org/10.1021/acs.molpharmaceut.0c00629

mutations that reduced viscosity.[19] Various other computational approaches have been used to study protein−excipient interactions,[20−22] including our previous studies on antibody−excipient interactions using molecular dynamics (MD) simulations.[9,23] However, simulations are often opaque as they do not directly yield a numerical relationship between chemical surface properties and excipient interactions. Instead, they require lengthy manual analysis of the results to draw conclusions about protein−excipient interaction mechanisms.[19,22,23]

Machine learning approaches can be used to analyze simulation results to look for trends in interactions. Although machine learning approaches have frequently been used for other antibody applications, including antibody structure prediction[24] and antigen binding,[25−27] they have not yet been applied specifically to model antibody−excipient interactions. Linear models will yield a direct numerical relationship between antibody surface properties and the local excipient concentration, although likely with less accuracy than nonlinear models, but it is more difficult to extract physical insight from nonlinear models. If reasonably high accuracy is achieved with a nonlinear model, antibody−excipient interactions could be predicted in a matter of seconds instead of days. Rapid prediction of antibody−excipient interactions, as well as an understanding of the impact of these interactions on physical stability of antibodies, is desired for early-stage formulation design. This would reduce the number of experiments required, reducing the overall cost of formulation development.

Using machine learning models requires numerically representing the antibody surface. Previously, various groups have worked to identify features that might be important in describing the proteins, including representations of amino acids,[28] molecular curvature,[29] and 3D Zernike descriptors for mAbs.[30] However, thus far, no systematic framework for representing the chemical properties of local regions of a mAbs surface has been developed. Additionally, to the best of our knowledge, no systematic studies of how antibody surface properties affect the interactions with a large panel of excipients exist beyond our previous work.[9,23]

Here, we use machine learning algorithms to connect antibody surface features to local preferential interactions with a variety of excipients. First, a feature set for numerically describing local surface regions of mAbs is developed. Machine learning models are trained using these features to predict local mAb−excipient preferential interactions as calculated via MD simulation. We demonstrate the use of this approach on formulation excipients sorbitol, sucrose, trehalose, proline, Arg·HCl, and NaCl, identifying the top surface characteristics that determine interactions. Finally, we demonstrate how this tool might be used in formulation design for novel antibodies via case studies.

■ MATERIALS AND METHODS

Simulation Methods. Molecular simulations were carried out according to our previously described methods.[9,23] Briefly, MD simulations were performed using GROMACS 5.0.5[31] and the CHARMM36m[32,33] force field, with the exception of force field parameters for the following excipients: sorbitol, sucrose, trehalose, proline, and Arg·HCl. Force field parameters for these excipients were taken from our previous work.[9,23,34] Simulations were performed at pH 5.5, with amino acid protonation states set using PropKa on the PDB2PQR server.[35] Simulations were made charge neutral with the addition of sodium or chloride ions. All results reported here are based on simulations with a bulk excipient concentration of 0.5 *m*.

Energy minimization and 15 ns of equilibration were performed prior to beginning production simulations, and configurations were saved every 10 ps. Simulations were run for 70 ns, with the exception of simulations with the excipient Arg·HCl, which were run for 200 ns. The longer simulation time for Arg·HCl ensured convergence despite Arg·HCl's more complicated clustering behaviors.[22] The Python package MDAnalysis[36,37] was used for trajectory analysis, and statistical errors were calculated using the methods of Allen and Tildesley.[38]

Preferential Interaction Coefficient Calculation. Overall preferential interaction coefficients were calculated from MD simulations according to the methods of Baynes and Trout[39] and Shukla et al.[40] For an entire protein, they can be calculated according to

$$\Gamma_{23}(r) = \left\langle n_3(r) - n_1(r)\left(\frac{n_3^{\text{total}} - n_3(r)}{n_1^{\text{total}} - n_1(r)}\right)\right\rangle \tag{1}$$

where the subscript 1 refers to water, 2 refers to the protein, and 3 refers to the excipient. $n_i(r)$ refers to the number of molecules of type *i* within a distance *r* of the protein van der Waals surface. Past some distance *R*, usually around 8 Å, the value of $\Gamma_{23}(r > R)$ converges to a constant value. It is this value, referred to as Γ_{23} that is compared to the experimental result.

Equation 1 can be adapted for calculating the local Γ_{23} value for each residue. When calculating local (per-residue) Γ_{23} values, the subscript 2 refers only to atoms of that particular protein residue. In order to avoid double-counting, each excipient and water atom is "assigned" to its closest protein residue and then the total number of excipient and water atoms assigned to each residue is converted back to the corresponding number of excipient and water molecules. This allows fractions of molecules to be assigned to each residue, which better describes the distribution of excipient molecules. These numbers are used to calculate the local Γ_{23}. As with the overall Γ_{23} value, we found that local Γ_{23} values tended to converge by a distance of 8 Å. Using this method, the sum of the local Γ_{23} equals the overall Γ_{23} value.

In the case of 1:1 ionic excipients, the local Γ_{23} values for each excipient are calculated separately and then combined through eq 2.

$$\Gamma_{23,\text{local}}(r) = \frac{\Gamma_{2,+} + \Gamma_{2,-}}{2} \tag{2}$$

where $\Gamma_{2,+}$ and $\Gamma_{2,-}$ refer to the Γ_{23} values for the cation and anion, respectively, calculated according to eq 1. Note that this slightly differs from the equation given by Record and Anderson to calculate the Γ_{23} value of the entire molecule.[41] Record and Anderson subtract the net charge of the molecule from the numerator of eq 2. Thus, the overall molecule's Γ_{23} value can be obtained by summing the local Γ_{23} values of all the residues in the molecule and then subtracting half the net charge magnitude.

Antibody Data Set. The machine learning models developed here used features of individual residues to predict local Γ_{23} values. For this work, a data set of approximately 40,000 antibody residues was developed. These residues were

Table 1. Features Used to Describe Each Antibody Residue[a]

category	feature	subtypes	description	NBH?
sequence	hydrophobicity		residue hydrophobicity	n
	residue identity		one-hot vector indicating residue identity (3 encodings for histidine)	n
structure	protrusion index (CX)[48]	mnimum	protrusion from mAb surface, calculated for each atom in a residue. Minimum is lowest value of any atom in the residue	y
		maximum	maximum CX of any atom in the residue	y
		average	average CX value of all atoms in the residue	y
	depth index (DPX)[49]	minimum	distance from the closest point on the mAb surface, calculated for each atom in a residue. Minimum is lowest value of any atom in the residue	y
		maximum	maximum DPX of any atom in the residue	y
		average	average DPX value of all atoms in the residue	y
	electrostatic potential[29]		a charge metric modified by the distance to the surface	y
	SASA	total	SASA of all atoms in the residue	y
		hydrophobic	SASA of atoms in Ala, Ile, Leu, Met, Phe, Trp, Tyr, and Val	y
		hydrophilic	SASA of atoms in Arg, His, Lys, Asp, Glu, Ser, Thr, Asn, and Gln	y
		positive	SASA of atoms in Arg, Hsp, and Lys	y
		negative	SASA of atoms in Asp and Glu	y
		backbone	SASA of backbone C, O, N, and attached H atoms	y
		sidechain	SASA of atoms in the R group	y
	net charge	total	residue net charge	y
		exposed	sum of partial charges of atoms with SASA > 0	y
	shape index[29]		description of concavity/convexity	[b]
	SAP[8]		measure of surface-exposed hydrophobicity of atoms within 5 Å	only 5 Å[c]
	SCM[7]		measure of exposed charge within 10 Å	only 10 Å[d]
	secondary structure		one-hot vector indicating secondary structure (α-helix,β-sheet, loop)	n
	fractional exposure		ratio of the residue's SASA to the standard exposure of the residue in Ala-X-Ala	n

[a] Category refers to the broad category for each feature (sequence-based or structural). Subtypes are listed for any related features. Features calculated both for the residue alone and for the residue plus neighboring residues within some cutoff distance (3, 5, 7, 10, 12, 15, 20, and 25 Å) are identified in the neighborhood (NBH) column. [b] The shape index was not calculated for the residue alone and neighborhood values were calculated at 3, 4, 5, 6, 7, 8, 9, and 10 Å. [c] The SAP tool was designed for calculations at 5 Å, as discussed by Chennamsetty et al.[8] [d] The SCM tool was designed for calculations at 10 Å, as discussed by Agrawal et al.[7]

from the IgG1 antibodies listed in Table S1, which indicates the antibody name, the PDB ID if appropriate, and the type of simulation (full antibody or Fab only). The antibodies labeled with numbers are those whose sequences were provided by AstraZeneca. Full antibody structures were generated according to the previously described methods.[9,23,42] Additionally, the entire data set is included in the Excel file in the Supporting Information. The order of the residues in this data set is shuffled to maintain anonymity of proprietary AstraZeneca mAbs.

Each antibody residue was described by a set of 182 features, leading to a data set of 43,816 residues × 182 features. The features used to describe each residue are an important result of this work and are described in more detail in the Results and Discussion section and are summarized in Table 1. Feature values were calculated from the energy-minimized structure of the antibody in water. Prior to training the machine learning models, features were rescaled and normalized to have a mean of 0 and a standard deviation of 1.

Machine Learning Models. Two machine learning models were considered: elastic net (EN) models and support vector machine (SVM) models. Both were implemented using the sklearn Python package.[43] An EN is a simple linear model which identifies the least-squares weights for each feature that yield the lowest error. EN models include both L1 and L2 regularization to force the weights of unimportant features to 0 and to avoid overfitting. The optimal weights, $\hat{\beta}$, are determined according to eq 3, where y refers to the target vector, X refers to the feature matrix, and λ_1 and λ_2 are regularization parameters. The EN model is fully described by the two λ parameters.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_1\|\beta\| + \lambda_2\|\beta\|^2) \tag{3}$$

We used SVM models with Gaussian radial basis functions. Berwick has made a clear explanation of SVM models.[44] The SVM model was implemented in Python using the sklearn package. The SVM model with Gaussian radial basis functions is fully described by the parameters C and γ.

Neural network models were also considered. Neural networks have been used for other antibody applications, including paratope prediction.[25,45,46] However, the neural networks that we tested to model the local Γ_{23} values did not perform better than the SVM models (data not shown). Because neural networks have many more fitted parameters compared to SVM models, there are more opportunities to overfit the data and the resulting models would be less robust. Thus, neural networks were not included in this study.

The data set used in developing these models consisted of the approximately 40,000 residues introduced in the previous section. These data were divided into three sets: training, validation, and testing. The testing set, used to assess the performance of the models, consisted of the data for mAbs A, B, C, and D, which account for about 10% of the total data. The data in the testing set were not included in model training. The remainder of the data made up the training and validation sets. The models were trained using fivefold cross validation for hyperparameter tuning.

Model performance was characterized based on the root-mean-squared error (RMSE), accuracy, and true positive rate (TPR). To calculate the accuracy and TPR, which are classification metrics, the regression results were first binned into "low-interaction" and "high-interaction" classes, with the class boundaries chosen based on the data distribution such that about half the data would be in each class. The "high-interaction" class was considered the "positive" class for accuracy and TPR calculation.

Viscosity Measurements. The viscosity was measured in each excipient using a DHR-3 (TA Instruments) cone and plate rheometer using a 1° cone with 20 mm diameter at 20 °C. The dynamic viscosity was determined for shear rates between 800 and 2000 s^{-1}. Measurements were performed in triplicate. Solution concentrations were verified using a NanoDrop One (Thermo Scientific). The buffer conditions for all experiments were 25 mM sodium acetate (NaOAc), pH 5.5. For each excipient, the viscosity of mAbD at 200 mg/mL was calculated using the viscosity curve generated by measuring the viscosity at four or more concentrations between 140 and 230 mg/mL. Experimental viscosities and aggregation rates (from size-exclusion chromatography) for mAbs A, B, and C, as well as the experimental methods, can be found in our previous papers.[9,16,17,23]

RESULTS AND DISCUSSION

Feature Vector. We sought to develop a machine learning model that predicted the local Γ_{23} value over the surface of an antibody. This required subdividing the antibody surface into small units, each with a separate local Γ_{23} value. These units needed to be small enough to provide insight into the local distribution of the excipient around the antibody surface but large enough that there could be meaningful antibody—excipient interactions within each unit. For example, the entire Fv domain is not an appropriate unit. Although it is defined for any antibody and thus is transferable, it is too large to capture the local behaviors that we have previously identified to be important.[9] Conversely, although an individual atom is also transferable, it is too small a unit to be relevant. Effectively, the unit should be on the order of the size of an excipient molecule. To fit these size criteria and to be easily applicable to any antibody, we chose individual residues as these small units.

Any subunit of the antibody surface has artificial boundaries, as the boundary of a residue has no impact on how the protein surface in that area interacts with an excipient. Thus, the feature set needed to include information about the local neighborhood, which are its neighboring residues. This was accomplished by calculating feature values not only for atoms in the residue but also for atoms within certain distances of any atom in the residue. For example, one feature describing a single residue is the solvent-accessible surface area (SASA) of the residue. Another feature describing the same residue is the SASA of that residue plus all atoms within 5 Å of that residue. This is an analogous idea to filtering used in image recognition or to junction tree encoding.[47]

A total of 182 features were used to describe each residue. Table 1 summarizes the features used. Some of these features were also calculated over the residue and its neighboring residues, as indicated in the neighborhood column. Unless otherwise noted, neighborhood feature values were calculated at distances of 3, 5, 7, 10, 12, 15, 20, and 25 Å.

Features were split into two broad categories: sequence-based and structure-based. Sequence-based features are those that can be obtained just from the amino acid sequence, such as the residue identity. Structure-based features are those obtained from the 3D structure of the antibody. These include features such as the SASA, the charge, the spatial aggregation propensity (SAP),[8] and the spatial charge map (SCM).[7]

It is not assumed that all these features must be relevant. Instead, this feature set was developed to include any chemical property that might be relevant. The training of the machine learning models will determine the appropriate weights for each feature, with unimportant features having lower weights.

Antibody Data Set. The data set used in this study consisted of the residues of the 41 antibodies listed in Table S1. Some structures were for full antibodies, while others were just for the Fab domains. We have found that the local Γ_{23} value residues in an antibody in a Fab-only simulation generally match those of the corresponding residues in a full antibody simulation, as the interactions we are examining are local. Thus, we could mix data from Fab-only and full antibody simulations. Only residues with nonzero surface exposure were included in the data set. This led to a data set of approximately 40,000 residues, which was randomly divided into training and testing sets.

In principle, this approach should be applicable to any protein. However, the data set contained only residues from antibody structures as that was the focus for this work. If this approach is to be extended to include other types of proteins, similar proteins should be included in the training data.

The target values are the local Γ_{23} values for each residue, calculated as described in the Materials and Methods section. The feature values were calculated from energy minimized structures. We compared the performance of models using just the energy-minimized structure to that of models trained with features averaged over a 50+ ns simulation of the mAbs in water and found that the RMSEs and accuracies were virtually identical. Because obtaining energy-minimized structures is much less time-consuming, as no MD simulation is required and there was no loss in performance, we chose to use features calculated from energy-minimized structures. Figure S1 also demonstrates visually that both these approaches lead to similar performance.

Machine Learning Models. Two types of machine learning models were considered: ENs and SVMs. Separate models were developed for each excipient included in this study: sorbitol, sucrose, trehalose, proline, Arg·HCl, and NaCl. Generally, the same hyperparameters were used for the carbohydrates, proline, and NaCl, while a different set of hyperparameters were used for Arg·HCl. The hyperparameters for the EN and SVM models developed are shown in Table 2. A visual example of the performance of the NaCl models on mAbC, an antibody excluded from the training and testing set, is shown in Figure 1.

Figure 1 shows how, while neither model is a perfect match for the simulation result, both capture much of the relative excipient distribution around the antibody. The RMSE, accuracy, and TPRs are reported in Table 3. The SVM

Table 2. Hyperparameters for Machine Learning Models[a]

	EN		SVM	
	λ_1	λ_2	C	γ
carbohydrates, proline, NaCl	0.01	0.0001	40	0.001
Arg·HCl	0.000316	4.64×10^{-7}	55	0.001259

[a] Carbohydrates refers to sorbitol, sucrose, and trehalose.

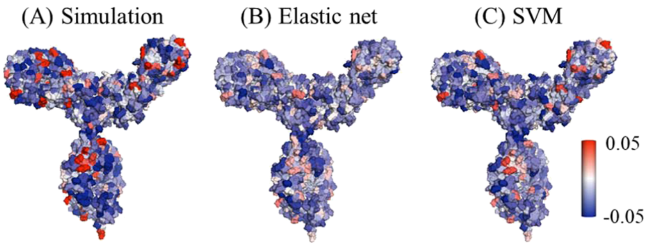


Figure 1. Views of mAbC interactions with NaCl, colored by the local Γ_{23} value (based on a bulk excipient concentration of 0.5 m). (A) Simulation result, which is the target of the machine learning models. (B,C) EN model and SVM model, respectively. Both clearly capture the trends in the distribution of NaCl around the antibody, with the SVM model performing slightly better. However, neither perfectly captures the magnitude of all the interactions.

models reduce error and improve accuracy as compared to the naïve baseline case, which predicts the average Γ_{23} value for all residues. In addition, the SVM models have significantly improved accuracies and TPRs over the EN models.

EN Feature Weights. Often, it can be challenging to gain biological insight from statistical learning algorithms.[50] One benefit of the EN, which is a linear model, is that it provides some insight into the features that are important to the model. A trained EN is a set of weights, one for each feature, with the weighted sum of the feature vector yielding the target value. Prior to training the EN, all features are rescaled and normalized, meaning that the values for each feature generally have an average of 0 and a standard deviation of 1. This normalization allows the weights to be compared across different features, as the possible magnitudes of each feature are approximately the same. Figure 2 shows, for each excipient, the features with weights of at least 4%. Note that Figure 2 shows weights with positive or negative values. Features with a positive weight contribute positively to the local Γ_{23} value, meaning higher values of that feature correspond to more local excipient molecules, while features with a negative weight contribute negatively, meaning higher values of that feature correspond to more local water molecules.

Interestingly, the hydrophilic SASA is the only feature with a contribution of at least 4% to the local Γ_{23} values for each type of excipient. For the carbohydrates and proline, it has a strong negative impact, while for NaCl, it has a weak negative impact and for Arg·HCl, it has a slightly positive impact. This suggests that the carbohydrates and proline tend to interact less with hydrophilic residues compared to Arg·HCl and NaCl, which makes sense given that Arg·HCl and NaCl are ionic compounds.

As shown in Figure 2, sorbitol, sucrose, and trehalose had similar high-weight features. These included the hydrophobic surface area, which had the highest positive weight for all three excipients, as well as the positive, negative, hydrophilic, and side chain surface areas, which had some of the highest negative weights. These similarities suggest that the three excipients interact through a similar mechanism, which makes sense, as they are all carbohydrates and expected to interact primarily through the alcohol groups. The positive contribution of the hydrophobic surface area agrees with our previous result indicating that the carbohydrate excipients tend to interact more strongly with hydrophobic regions.[9] The negative contributions of the hydrophilic and charged surface areas to the local Γ_{23} indicate that these hydrophilic residues tend to interact with water rather than the carbohydrate excipients.

Interestingly, many of the top features for proline also appear among the top features for the carbohydrate excipients. For example, the hydrophobic SASA has the highest positive weight, and the positive and hydrophilic SASAs have some of the most negative weights. As with the carbohydrates, this suggests that proline interacts more strongly with hydrophobic residues, while water interacts more strongly with hydrophilic residues. Previous studies have shown that the pyrrolidine side chain of proline interacts with hydrophobic residues via CH−π interactions.[51,52]

For Arg·HCl, many of the top features are residue names. These features are from the one-hot residue name feature set, in which the one high value indicates the residue identity. Specifically, all five charged residues appear in the top five features for Arg·HCl. The three positively charged residues have positive weights, while the two negative residues have negative weights. However, the net charge of the residue has a negative weight, leading to a positive contribution of the negatively charged residues to the local Γ_{23} values. Because the local Γ_{23} value is based on all the features in the feature set, there can be several competing effects. The Arg·HCl plot does show that features related to charge are important and may provide further insight into these competing effects.

Table 3. RMSEs, Accuracies, and TPRs for the Naive Baseline Model, EN, and SVM Models for Each Excipient[a]

	naïve baseline			EN			SVM		
excipient	RMSE	Acc.	TPR	RMSE	Acc.	TPR	RMSE	Acc.	TPR
sorbitol	0.032	0.45	0	0.017	0.85	0.92	0.017	0.85	0.90
sucrose	0.041	0.45	0	0.023	0.85	0.87	0.021	0.86	0.89
trehalose	0.038	0.45	0	0.021	0.85	0.88	0.019	0.86	0.89
proline	0.024	0.54	0	0.017	0.81	0.89	0.016	0.83	0.86
Arg·HCl	0.092	0.57	0	0.086	0.65	0.61	0.083	0.78	0.80
NaCl	0.025	0.50	0	0.019	0.79	0.84	0.017	0.82	0.85

[a] The naïve baseline predicts the average Γ_{23} value for each residue. Results shown for performances on the test set. For all excipients, the SVM models have slightly lower error and higher accuracies compared to the EN models.

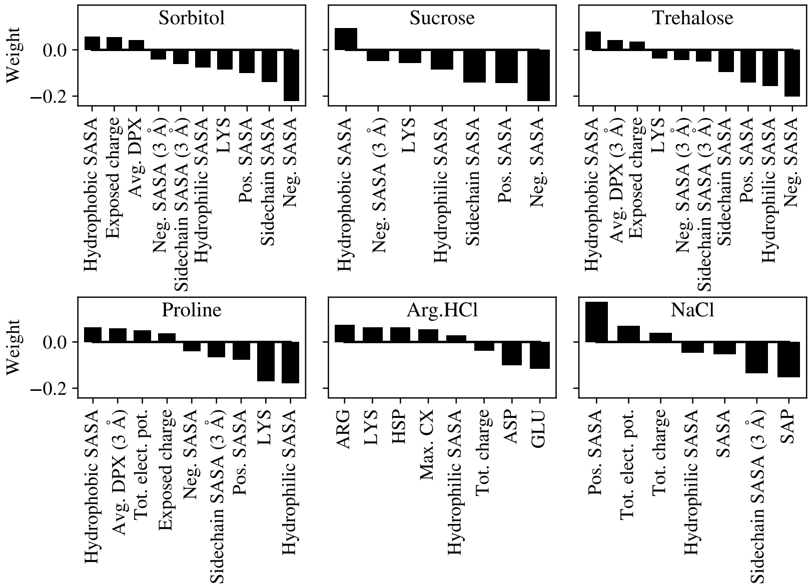


Figure 2. Plots of the feature weights from the ENs for features with weights above 4%. Positive values indicate a positive contribution to the local Γ_{23} value, and negative values indicate a negative contribution. Feature explanations can be found in Table 1. In general, similar features show up as the top features for sorbitol, sucrose, and trehalose, indicating that the underlying mechanism of interaction of these excipients with the antibody is the same. Interestingly, there is some overlap with the top features for proline interactions, indicating some similarities in how these osmolytes interact. For Arg·HCl, the identity of the surface residue tends to have a high impact, as indicated by many of the top features being a particular type of residue. For NaCl, charge features tend to be most important.

For NaCl, the top features are mostly related to charge or hydrophobicity. It makes sense that the feature with the highest positive weight is the positively charged SASA because NaCl is expected to interact primarily through electrostatic interactions. The feature with the most negative weight, the SAP, is effectively a measure of surface hydrophobicity. This suggests that more hydrophobic regions on the surface interact more strongly with water than with NaCl.

Case Studies for Formulation Design. Previous studies[9,23] have identified important antibody−excipient interactions that are thought to be related to the observed experimental behavior. Here, we examine if those interactions are captured using the SVM model and describe how that insight might be used in formulation design. Then, we demonstrate how we applied this approach to an antibody prior to assessing its viscosity experimentally. In all the case studies presented here, the antibodies were excluded from the training and testing sets. Thus, the SVM results presented represent the model performance on an antibody entirely not previously encountered using the model.

Aggregation. Our previous studies on mAbA indicated that at excipient concentrations of 100 mM, Arg·HCl and NaCl increased the aggregation rate by factors of about 1.6 and 2.3, respectively, while proline had no effect.[23] At 200 mM, the carbohydrates reduced aggregation.[9] Because only the ionic excipients increased aggregation, these increases are thought to be related to charge effects. For mAbA, these behaviors were examined by considering the net local Γ_{23} values of the Fv domain as a whole, as well as of the positively and negatively charged residues within the Fv domain. Figure 3 compares these values as calculated from the simulations (A) and as calculated based on the SVM models (B).

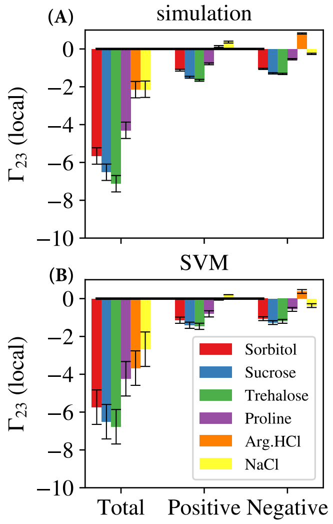


Figure 3. Sums of the Γ_{23} values in the variable region of mAbA from (A) simulations with each excipient or (B) SVM model outputs, based on a bulk excipient concentration of 0.5 *m*. The results are further broken down by summing over just the positively or negatively charged residues in the Fv. The SVM models generally capture the appropriate magnitudes of the total Γ_{23} values, although the SVM model for Arg·HCl tends to predict more exclusion. The SVM model does capture that Arg·HCl is included near negative residues, while NaCl is excluded. This behavior may be related to why NaCl increases mAbA aggregation more than Arg·HCl. Error bars in (A) are the statistical error of the simulation, calculated according to Allen and Tildesley.[38] Error bars in (B) are calculated from the RMSE for the SVM for individual residues in the Fv domain.

Near positive residues, the simulations (Figure 3A) indicate slight inclusion of Arg·HCl and NaCl, while near negative residues, they indicate inclusion of Arg·HCl but exclusion of NaCl. As shown in Figure 3B, the SVM models capture all these behaviors within the error tolerance. Our previous study suggested that the increase in aggregation of mAbA in Arg·HCl and NaCl was likely due to a reduction in electrostatic repulsion in the presence of these excipients, as mAbA has

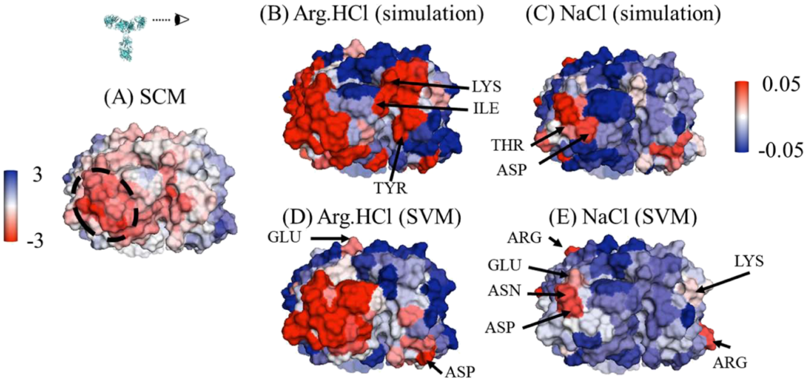


Figure 4. Views of the mAbC variable region. (A) Fv colored by the SCM score, with red indicating the exposed negative charge (left color bar). (B−E) Fv colored by the local Γ_{23} value (right color bar). (B,C) results from simulations with the excipients Arg·HCl or NaCl, respectively, while (D,E) the predicted values using the SVM models, all based on a bulk excipient concentration of 0.5 *m*. Both the SVM model and the simulation result indicate that these excipients are net-included near the lowest-SCM region circled in (A).

strong positive charges on the Fv and Fc at pH 5.5.[23] The weaker interaction of NaCl with negative residues may contribute to the greater increase in the aggregation rate in the presence of NaCl (increase of 138%) compared to Arg·HCl (increase of 57%) and this result is captured using the SVM models. Future studies at other pH values would allow for a more full comparison between Arg·HCl and NaCl and understanding of why NaCl interacts less strongly with negative residues.

We performed a similar analysis to understand the impact of trehalose on the aggregation of mAbB. Our previous study found that trehalose reduced the monomer loss rate of mAbB by about 32%,[17] and analysis of local excipient interactions suggested that this was due to interactions of trehalose with hydrophobic residues.[9] We confirmed that this behavior was captured using the SVM models. Both the simulation and the SVM model indicate that the sum of the local Γ_{23} value for trehalose near hydrophobic residues in the Fv was −0.8. Additionally, both indicate that the sum of the local Γ_{23} values near aromatic residues in the Fv was −0.3. In comparison, if trehalose was excluded, the average amount from these types of residues and the local exclusions would be −3.3 and −2.8, respectively. This result, captured using the SVM models, indicates that trehalose is much less excluded than the average from hydrophobic and aromatic residues in the Fv and these interactions could lead to the observed disruption of aggregation.

Viscosity. Our previous study on mAbC found that its viscosity at 100 mg/mL reduced from about 35 cP at pH 5.5 to about 5 cP with the addition of 100 mM Arg·HCl or NaCl.[23] The high viscosity of mAbC is likely related to the negatively charged patches in the Fv, which may also contribute to the propensity for reversible self-association.[53] These negative charge patches are shown in red in Figure 4A. Figure 4B,C shows the local Γ_{23} values for Arg·HCl and NaCl calculated from simulation, while Figure 4D,E shows the local Γ_{23} values output using the SVM models. Although the simulation and SVM results are not identical, the SVM results do capture the important excipient interactions with the circled high-SCM region that are expected to be responsible for the observed reduction in viscosity.

The simulation results indicate net inclusion of Arg·HCl (local Γ_{23} = 2.9) and NaCl (local Γ_{23} = 0.2) with the highest-SCM region circled in Figure 4A. The strong interactions with this negatively charged region are expected to disrupt mAbC−mAbC self-interactions that lead to increased viscosity. The SVM results also indicate net inclusion of Arg·HCl (1.6) and NaCl (0.02), although both are lower in magnitude than the corresponding simulation results. Even so, the SVM models still capture the much stronger than average interactions in the negatively charged region, which is necessary for implementing the SVM models in formulation design. For example, if a proposed antibody were found to have a large amount of negative charge in the Fv domain, such as mAbC, the SVM model could be used to check if ionic excipients will interact strongly in that region. If so, the results with mAbC suggest that the viscosity in those excipients will be reduced to acceptable levels.

Demonstration of the Approach: mAbD Viscosity. For the previous case studies, the experimental work was carried out prior to the simulation work. In order to perform an evaluation on an antibody which we had not previously studied, we evaluated a new antibody, mAbD, first through simulation and then through experiment. We began using the SAP[8] and SCM[7] tools to characterize the surface of the Fv domain. The low overall SCM score, 375 ± 22, and the high net positive charge overall, +36, and on the Fv, +9, suggested that the antibody would have low viscosity at 100 mg/mL. There is a small region of exposed negative charge in the Fv domain, but mostly exposed positive charge, which is expected to lead to electrostatic repulsion and thus low viscosity. The low SCM score and the relatively small amount of exposed negative charge in the variable region suggest that charge effects do not dominate the viscosity behavior of mAbD.

Based on this analysis of the mAb Fv surface and our experience with mAbs A, B, and C, we expected carbohydrates to have little effect on the viscosity or to increase the viscosity. This prediction is based on He et al.'s results[54] that sugars tended to increase the viscosity of antibodies at high concentrations, with larger sugars leading to greater increases in viscosity. This increase, which was significantly more than the increase in viscosity of the buffer alone with the addition of sugars, was attributed to greater preferential exclusion of larger carbohydrates.[55] In addition, He's results generally agreed with our previous results for the viscosity behavior of mAbs B and C in carbohydrates,[9] although there were some exceptions based on specific self-association behaviors. These previous results are reproduced in Figure S2 in the Supporting Information.

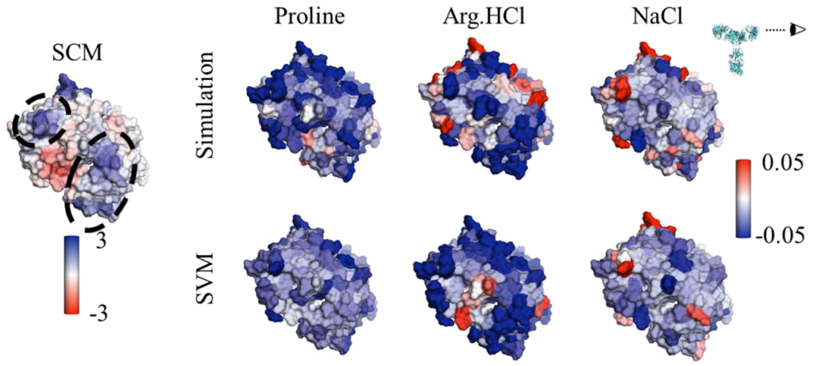


Figure 5. Views of the mAbD variable region colored by the SCM score (far left) or by the local Γ_{23} values with each excipient. The top row of plots is colored by the local Γ_{23} values based on simulations performed with bulk excipient concentrations of 0.5 *m*, while the bottom row of plots is colored by the local Γ_{23} values predicted using the SVM model. Both the simulation and SVM models indicate that proline and Arg·HCl are excluded from the circled positive charge regions, while the net Γ_{23} value for NaCl near these regions is around 0. This suggests that NaCl interferes more with the electrostatic repulsion of mAbD from other mAbD molecules at pH 5.5, leading to the observed increase in viscosity compared to behavior in the buffer.

We expected slightly more complicated behaviors in the presence of proline, Arg·HCl, and NaCl. From our previous work with these excipients,[23] we knew that they can impact viscosity through interactions with charged residues and hydrophobic residues. Figure 5 shows the local interaction maps of these excipients with the variable region of mAbD. In addition, it shows the SCM scores mapped onto the variable region, with the two positively charged regions circled.

Similarly to mAbA, we expected that interactions with these positively charged regions might disrupt the electrostatic repulsion between molecules, leading to higher viscosity. According to the SVM models of the mAbD behavior in 0.5 *m* excipient, proline and Arg·HCl are fairly strongly excluded from these regions, with local Γ_{23} values of −0.51 and −1.04, respectively. Importantly, these values are both more negative than would be expected if the excipients were excluded the average amount from these residues. In general, exclusion from this region suggests that the electrostatic repulsion between mAbD molecules will be maintained, leading to the same or slightly lower viscosity than in buffer.

The SVM model indicates slight exclusion of NaCl from this region, with a local Γ_{23} value of −0.08. However, this value is higher than what would be expected if NaCl was excluded from each residue in this patch by the average local Γ_{23} value. This average exclusion would correspond to a local Γ_{23} value of −0.29. Thus, NaCl has above-average interactions with this positively charged region of mAbD. This above-average interaction was expected to correspond to a slight increase in the viscosity of mAbD in NaCl because it could disrupt the electrostatic repulsion between mAbD molecules.

We compared the SVM results with the simulation results. The simulation results also indicated strong exclusion of proline (Γ_{23} = −0.7) and Arg·HCl (Γ_{23} = −1.2) from the circled high-SCM regions. The simulation indicated slight inclusion (Γ_{23} = 0.1) of NaCl with these regions. This agrees with the SVM results which found that proline and Arg·HCl are more excluded than average, while NaCl is more included than average.

Finally, we experimentally validated these results by measuring the viscosity of mAbD in buffer and in the presence of each excipient. As we expected, mAbD had a low viscosity, around 10 cP, at 100 mg/mL in 25 mM sodium acetate buffer, pH 5.5. In order to observe and differentiate between the impacts of the different excipients, we measured the viscosities at higher concentrations of the antibody. Figure 6 shows the viscosity behavior in each excipient at 200 mg/mL.

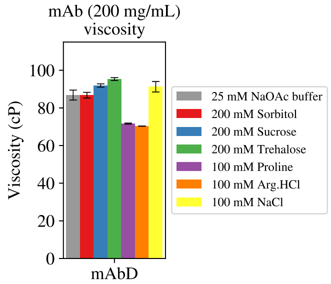


Figure 6. mAbD viscosity at 200 mg/mL at pH 5.5. Sucrose, trehalose, and NaCl increase the viscosity, compared to the viscosity in the buffer, while proline and Arg·HCl reduce the viscosity.

Figure 6 confirms our predictions for the viscosity behavior of mAbD in each of the excipients. Of the three carbohydrates, sorbitol has no effect on the viscosity while sucrose and trehalose increase the viscosity. Proline and Arg·HCl reduce the viscosity, while NaCl increases the viscosity. This demonstrates that the SVM models were able to capture the key interactions that might be responsible for the observed viscosity behavior. It is important to note that based on our experience, it is necessary to observe the properties of the mAb surface, such as SAP and SCM, to understand the potential impact of each excipient. The surface characteristics will influence the underlying aggregation and viscosity mechanisms, which in turn are important for understanding how these behaviors will change in the presence of the excipient.

The results described here have focused on studying individual excipients, while typical antibody formulations contain multiple excipients. Proper modeling of antibodies in multiexcipient mixtures may require considering excipient—excipient interactions. This study did not examine if excipients that interact via very different mechanisms, such as one by electrostatics and one by excluded volume effects, would have combined behaviors that are fairly similar to the sum of the individual interactions and this requires further study of

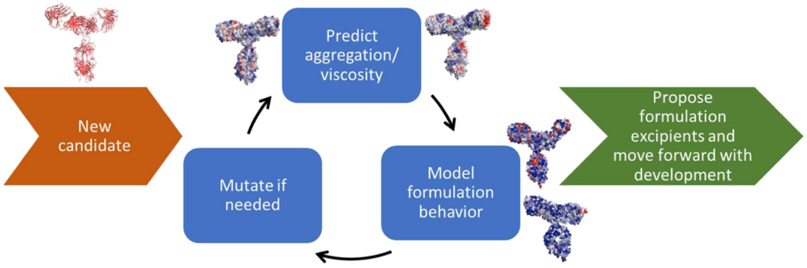

New candidate
Predict aggregation/ viscosity
Model formulation behavior
Mutate if needed
Propose formulation excipients and move forward with development

Figure 7. Proposed computational formulation design workflow.

multiexcipient mixtures. Excipients that interact via the same mechanism would likely have significant excipient−excipient interactions. However, if the excipients interact via identical mechanisms, it is possible that the combination would have the same overall impact on antibody behavior, if the local interactions are effectively the same. Future studies can examine how to extend this framework to multiexcipient mixtures.

Formulation Design Scheme. Based on these studies of antibody−excipient interactions, we propose the following formulation design scheme, also described in Figure 7:

1. candidate mAb sequence is converted to a structure using homology modeling
2. Surface characteristics, such as SAP and SCM, are modeled to identify problem patches which might contribute to elevated aggregation or viscosity
3. The SVM models for local Γ_{23} values presented here are used to predict local interactions with excipients
4. The local Γ_{23} sums near potentially problematic patches, such as those with exposed positive or negative charge, are used to predict the effect of the excipient on the mAb's stability
5. If there are expected to be aggregation or viscosity behaviors not addressed by the addition of excipients, point mutations can be performed and this process repeated from step 1
6. If aggregation and viscosity behaviors are expected to be addressed in the presence of certain excipients, the mAb structure and the stabilizing excipients move forward in the design and development process

This process can be applied to any new candidate antibody. Step 4 describes the most complicated stage, which involves analyzing the local excipient interactions in light of the antibody surface characteristics to predict the effect. Each of the case studies presented previously shows an example of this analysis being carried out and can be used as a guide in performing this analysis. The final outputs of this analysis are an optimized mAb sequence and the excipients expected to have a benefit in terms of propensity for aggregation and viscosity behavior.

CONCLUSIONS

In conclusion, we have developed a feature set that can be used to numerically represent the characteristics of small regions of an antibody's surface and demonstrated how these features can be used to model local antibody−excipient preferential interaction coefficients. The feature set contains both sequence information and structure information, such as charge, surface area, and hydrophobicity. We demonstrated that these features could be calculated for the energy-minimized antibody structure and used to model the local Γ_{23} values of sorbitol, sucrose, trehalose, proline, Arg·HCl, and NaCl with up to 86% accuracy using SVM models. In case studies with mAbs A, B, and C, the results of these models capture the relative strengths and weaknesses of the interactions of these excipients with regions of the mAb surface that are thought to be related to aggregation and viscosity behavior. This approach was validated by applying it to predict the impact of these excipients on the viscosity of a new antibody, mAbD, prior to carrying out experimental work. Finally, we presented a framework for computational formulation design using machine learning SVM models that have the potential to predict local antibody−excipient interactions to enable a mechanistic approach to directional formulation design. This framework, designed to be used in concert with existing mAb surface tools such as SAP and SCM, allows for the quantitative prediction of local antibody−excipient interactions, and the case studies included here provide guides for how to apply the results for formulation design.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.0c00629.

Data set used to train the machine learning algorithms (XLSX)

Additional detail on the performance of the EN versus the SVM, viscosity data for mAbs A, B, and C in the presence of carbohydrates, and detail on the data set (PDF)

AUTHOR INFORMATION

Corresponding Author

Bernhardt L. Trout − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*; orcid.org/0000-0003-1417-9470; Phone: 617-258-5021; Email: trout@mit.edu

Authors

Theresa K. Cloutier − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*; orcid.org/0000-0001-5964-9849

Chaitanya Sudrik − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Neil Mody − *Dosage Form Design and Development, AstraZeneca, Gaithersburg, Maryland 20878, United States*; orcid.org/0000-0003-2260-7658

Hasige A. Sathish − *Dosage Form Design and Development, AstraZeneca, Gaithersburg, Maryland 20878, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.molpharmaceut.0c00629

Notes

The authors declare the following competing financial interest(s): Neil Mody and Sathish Hasige are full time employees of AstraZeneca, which funded this research.

■ ACKNOWLEDGMENTS

This work was funded by AstraZeneca.

■ REFERENCES

(1) Kamerzell, T. J.; Esfandiary, R.; Joshi, S. B.; Middaugh, C. R.; Volkin, D. B. Protein-excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Adv. Drug Delivery Rev.* **2011**, *63*, 1118−1159.

(2) Daugherty, A. L.; Mrsny, R. J. Formulation and delivery issues for monoclonal antibody therapeutics. *Adv. Drug Delivery Rev.* **2006**, *58*, 686−706.

(3) Hofmann, M.; Gieseler, H. Predictive Screening Tools Used in High-Concentration Protein Formulation Development. *J. Pharm. Sci.* **2018**, *107*, 772−777.

(4) Muralidhara, B. K.; Wong, M. Critical considerations in the formulation development of parenteral biologic drugs. *Drug Discov. Today* **2020**, *25*, 574−581.

(5) Li, L.; Kumar, S.; Buck, P. M.; Burns, C.; Lavoie, J.; Singh, S. K.; Warne, N. W.; Nichols, P.; Luksha, N.; Boardman, D.; Vi, G. B. Concentration dependent viscosity of monoclonal antibody solutions: Explaining experimental behavior in terms of molecular properties. *Pharm. Res.* **2014**, *31*, 3161−3178.

(6) Tomar, D. S.; Kumar, S.; Singh, S. K.; Goswami, S.; Li, L. Molecular basis of high viscosity in concentrated antibody solutions: Strategies for high concentration drug product development. *mAbs* **2016**, *8*, 216−228.

(7) 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*, 43−48.

(8) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11937−11942.

(9) Cloutier, T.; Sudrik, C.; Mody, N.; Sathish, H. A.; Trout, B. L. Molecular computations of preferential interaction coefficients of IgG1 monoclonal antibodies with sorbitol, sucrose, and trehalose and the impact of these excipients on aggregation and viscosity. *Mol. Pharmacol.* **2019**, *16*, 3657−3664.

(10) Wang, S.; Zhang, N.; Hu, T.; Dai, W.; Feng, X.; Zhang, X.; Qian, F. Viscosity-Lowering Effect of Amino Acids and Salts on Highly Concentrated Solutions of Two IgG1 Monoclonal Antibodies. *Mol. Pharm.* **2015**, *12*, 4478−4487.

(11) Thakkar, S. V.; Joshi, S. B.; Jones, M. E.; Sathish, H. A.; Bishop, S. M.; Volkin, D. B.; Russell Middaugh, C. Excipients differentially influence the conformational stability and pretransition dynamics of two IgG1 monoclonal antibodies. *J. Pharm. Sci.* **2012**, *101*, 3062−3077.

(12) Lee, J. C.; Timasheff, S. N. The Stabilization of Proteins by Sucrose. *J. Biol. Chem.* **1981**, *256*, 7193−7201.

(13) Xie, G.; Timasheff, S. N. Mechanism of the Stabilization of Ribonuclease A by Sorbitol: Preferential Hydration is Greater for the Denatured than for the Native Protein. *Protein Sci.* **1997**, *6*, 211−221.

(14) Gekko, K.; Timasheff, S. N. Mechanism of Protein Stabilization by Glycerol: Preferential Hydration in Glycerol-Water Mixtures. *Biochemistry* **1981**, *20*, 4667−4676.

(15) Timasheff, S. N. The Control of Protein Stability and Association by Weak Interactions with Water: How Do Solvents Affect These Processes? *Annu. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 67−97.

(16) Sudrik, C.; Cloutier, T.; Pham, P.; Samra, H. S.; Trout, B. L. Preferential interactions of trehalose, L-arginine.HCl and sodium chloride with therapeutically relevant IgG1 monoclonal antibodies. *mAbs* **2017**, *9*, 1155−1168.

(17) Sudrik, C. M.; Cloutier, T.; Mody, N.; Sathish, H. A.; Trout, B. L. Understanding the Role of Preferential exclusion of sugars and polyols from native state IgG1 and its effect on monoclonal antibody aggregation and reversible self-association. *Pharm. Res.* **2019**, *36*, 109.

(18) Manikwar, P.; Majumdar, R.; Hickey, J. M.; Thakkar, S. V.; Samra, H. S.; Sathish, H. A.; Bishop, S. M.; Middaugh, C. R.; Weis, D. D.; Volkin, D. B. Correlating excipient effects on conformational and storage stability of an IgG1 monoclonal antibody with local dynamics as measured by hydrogen/deuterium-exchange mass spectrometry. *J. Pharm. Sci.* **2013**, *102*, 2136−2151.

(19) Tilegenova, C.; Izadi, S.; Yin, J.; Huang, C. S.; Wu, J.; Ellerman, D.; Hymowitz, S. G.; Walters, B.; Salisbury, C.; Carter, P. J. Dissecting the molecular basis of high viscosity of monospecific and bispecific IgG antibodies. *mAbs* **2020**, *12*, 1692764.

(20) Barata, T. S.; Zhang, C.; Dalby, P. A.; Brocchini, S.; Zloh, M. Identification of protein-excipient interaction hotspots using computational approaches. *Int. J. Mol. Sci.* **2016**, *17*, 853.

(21) Tosstorff, A.; Peters, G. H.; Winter, G. Study of the interaction between a novel, protein stabilizing dipeptide and Interferon-alpha-2a by construction of a Markov State Model from Molecular Dynamics simulations. *Eur. J. Pharm. Biopharm.* **2020**, *149*, 105.

(22) Shukla, D.; Trout, B. L. Interaction of Arginine with Proteins and the Mechanism by Which It Inhibits Aggregation. *J. Phys. Chem. B* **2010**, *114*, 13426−13438.

(23) Cloutier, T.; Sudrik, C.; Mody, N.; Sathish, H. A.; Trout, B. L. Molecular computations of preferential interactions of NaCl, arginine.HCl, and proline with IgG1 antibodies and their impact on aggregation and viscosity. **2020**, submitted.

(24) Norman, R. A.; Ambrosetti, F.; Bonvin, A. M. J. J.; Colwell, L. J.; Kelm, S.; Kumar, S.; Krawczyk, K. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings Bioinf.* **2019**, bbz095.

(25) Liberis, E.; Veličković, P.; Sormanni, P.; Vendruscolo, M.; Liò, P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* **2018**, *34*, 2944−2950.

(26) Liu, S.; Liu, C.; Deng, L. Machine learning approaches for protein-protein interaction hot spot prediction: Progress and comparative assessment. *Molecules* **2018**, *23*, 2535.

(27) Krawczyk, K.; Baker, T.; Shi, J.; Deane, C. M. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng., Des. Sel.* **2013**, *26*, 621−629.

(28) Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.* **2001**, *7*, 360−369.

(29) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comput. Chem.* **2002**, *23*, 1176−1187.

(30) Daberdaku, S.; Ferrari, C. Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics* **2019**, *35*, 1870−1876.

(31) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations Through Multi-level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(32) 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*, 71−73.

(33) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-

Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, ψ and Side-Chain $\chi1$ and $\chi2$ Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.

(34) Cloutier, T.; Sudrik, C.; Sathish, H. A.; Trout, B. L. Kirkwood-Buff-Derived Alcohol Parameters for Aqueous Carbohydrates and Their Application to Preferential Interaction Coefficient Calculations of Proteins. *J. Phys. Chem. B* **2018**, *122*, 9350−9360.

(35) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.

(36) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(37) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference Austin, Texas*, 2016; pp 102−109.

(38) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987; pp 191−195.

(39) Baynes, B. M.; Trout, B. L. Proteins in Mixed Solvents: A Molecular-Level Perspective. *J. Phys. Chem. B* **2003**, *107*, 14058−14067.

(40) Shukla, D.; Shinde, C.; Trout, B. L. Molecular Computations of Preferential Interaction Coefficients of Proteins. *J. Phys. Chem. B* **2009**, *113*, 12546−12554.

(41) Record, M. T.; Anderson, C. F. Interpretation of Preferential Interaction Coefficients of Nonelectrolytes and of Electrolyte Ions in Terms of a Two-Domain Model. *Biophys. J.* **1995**, *68*, 786−794.

(42) 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*, 905−913.

(43) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(44) Berwick, R. An Idiot's Guide to Support Vector Machines (SVMs), 2003 http://web.mit.edu/6.034/wwwbob/svm.pdf (accessed March 6, 2020).

(45) Chen, H.; Zhou, H.-X. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins* **2005**, *61*, 21−35.

(46) Kuroda, D.; Shirai, H.; Jacobson, M. P.; Nakamura, H. Computer-aided antibody design. *Protein Eng., Des. Sel.* **2012**, *25*, 507−521.

(47) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *35th International Conference on Machine Learning ICML 2018*, 2018; Vol. *5*, pp 3632−3648.

(48) Pintar, A.; Carugo, O.; Pongor, S. CX, an Algorithm that Identifies Protruding Atoms in Proteins. *Bioinformatics* **2002**, *18*, 980−984.

(49) Pintar, A.; Carugo, O.; Pongor, S. DPX: For the Analysis of the Protein Core. *Bioinformatics* **2003**, *19*, 313−314.

(50) Winslow, R. L.; Trayanova, N.; Geman, D.; Miller, M. I. Computational medicine: Translating models to clinical care. *Sci. Transl. Med.* **2012**, *4*, 158rv11.

(51) Hung, J. J.; Dear, B. J.; Dinin, A. K.; Borwankar, A. U.; Mehta, S. K.; Truskett, T. T.; Johnston, K. P. Improving viscosity and stability of a highly concentrated monoclonal antibody solution with concentrated proline. *Pharm. Res.* **2018**, *35*, 133.

(52) Zondlo, N. J. Aromatic-proline interactions: Electronically tunable CH/π interactions. *Acc. Chem. Res.* **2013**, *46*, 1039−1049.

(53) Arora, J.; Hu, Y.; Esfandiary, R.; Sathish, H. A.; Bishop, S. M.; Joshi, S. B.; Middaugh, C. R.; Volkin, D. B.; Weis, D. D. Charge-mediated Fab-Fc interactions in an IgG1 antibody induce reversible self-association, cluster formation, and elevated viscosity. *mAbs* **2016**, *8*, 1561−1574.

(54) He, F.; Woods, C. E.; Litowski, J. R.; Roschen, L. A.; Gadgil, H. S.; Razinkov, V. I.; Kerwin, B. A. Effect of sugar molecules on the viscosity of high concentration monoclonal antibody solutions. *Pharm. Res.* **2011**, *28*, 1552−1560.

(55) Sola-Penna, M.; Meyer-Fernandes, J. R. Stabilization Against Thermal Inactivation Promoted by Sugars on Enzyme Structure and Function: Why Is Trehalose More Effective Than Other Sugars? *Arch. Biochem. Biophys.* **1998**, *360*, 10−14.