

# Computational Characterization of Antibody–Excipient Interactions for Rational Excipient Selection Using the Site Identification by Ligand Competitive Saturation-Biologics Approach

Sunhwan Jo, Amy Xu, Joseph E. Curtis, Sandeep Somani, and Alexander D. MacKerell, Jr.\*



Cite This: *Mol. Pharmaceutics* 2020, 17, 4323–4333



Read Online

ACCESS |



Metrics & More



Article Recommendations



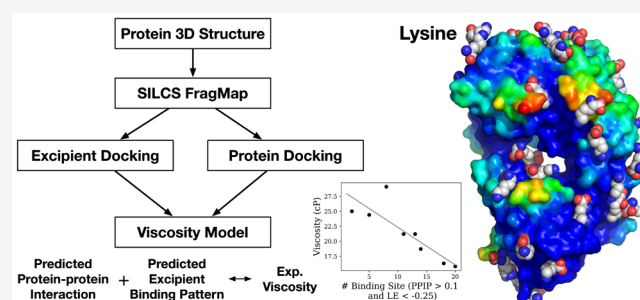
Supporting Information

**ABSTRACT:** Protein therapeutics typically require a concentrated protein formulation, which can lead to self-association and/or high viscosity due to protein–protein interaction (PPI). Excipients are often added to improve stability, bioavailability, and manufacturability of the protein therapeutics, but the selection of excipients often relies on trial and error. Therefore, understanding the excipient–protein interaction and its effect on non-specific PPI is important for rational selection of formulation development. In this study, we validate a general workflow based on the site identification by ligand competitive saturation (SILCS) technology, termed SILCS-Biologics, that can be applied to protein therapeutics for rational excipient selection. The National Institute of Standards and Technology monoclonal antibody (NISTmAb) reference along with the CNT0607 mAb is used as model antibody proteins to examine PPIs, and NISTmAb was used to further examine excipient–protein interactions, *in silico*. Metrics from SILCS include the distribution and predicted affinity of excipients, buffer interactions with the NISTmAb Fab, and the relation of the interactions to predicted PPI. Comparison with a range of experimental data showed multiple SILCS metrics to be predictive. Specifically, the number of favorable sites to which an excipient binds and the number of sites to which an excipient binds that are involved in predicted PPIs correlate with the experimentally determined viscosity. In addition, a combination of the number of binding sites and the predicted binding affinity is indicated to be predictive of relative protein stability. Comparison of arginine, trehalose, and sucrose, all of which give the highest viscosity in combination with analysis of  $B_{22}$  and  $k_D$  and the SILCS metrics, indicates that higher viscosities are associated with a low number of predicted binding sites, with lower binding affinity of arginine leading to its anomalously high impact on viscosity. The present study indicates the potential for the SILCS-Biologics approach to be of utility in the rational design of excipients during biologics formulation.

**KEYWORDS:** monoclonal antibody, molecular dynamics, biologics, protein-based drugs, formulation, protein–protein interactions

## INTRODUCTION

Protein therapeutics have been playing an increasing role as new therapeutic agents since the adoption of recombinant human insulin.<sup>1–3</sup> In fact, protein-based therapeutics represent about 30% of the newly approved drugs in 2016.<sup>4</sup> Formulation represents one of the key bottlenecks in the development of protein therapeutics because complex interactions of the proteins can lead to aggregation and high viscosity, which ultimately reduces the sample quality.<sup>5</sup> In addition, they must be formulated to maintain active protein levels during long-term storage and subsequent injection/infusion.<sup>6–8</sup> Therefore, formulation of protein therapeutics can be considered as an optimization process identifying combinations of excipients and buffers that act at the molecular level (1) to stabilize the native (folded) protein structure, preventing denaturation (unfolding) leading to loss of biological activity and non-specific aggregation of denatured protein molecules and (2) to minimize protein–protein interactions (PPI) of the native



structure, thereby avoiding aggregation and minimizing viscosity.<sup>9–11</sup>

Protein formulation invariably involves expensive and time-consuming experiments. To assist the formulation process, several computational tools have been developed. Many of these adopt bioinformatics approaches to predict aggregation-prone regions (APR) in a protein sequence.<sup>12</sup> However, many of the sequences used for training those bioinformatics algorithms are based on  $\beta$ -sheet forming sequences, such as amyloidogenic sequences,<sup>13–16</sup> and the utility of applying such

**Received:** July 25, 2020

**Revised:** September 22, 2020

**Accepted:** September 23, 2020

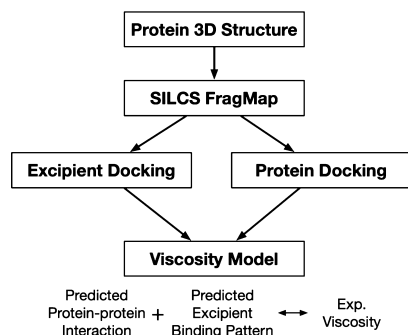
**Published:** September 23, 2020



algorithms to predict aggregation of globular proteins might be uncertain. Structure-based methods to predict aggregation have also been developed. The spatial aggregation propensity (SAP) score has shown that the higher SAP score, which indicates that the protein has larger exposed hydrophobic surfaces during molecular dynamics (MD) simulation, is associated with increased protein aggregation.<sup>17</sup> Similarly, spatial charge map was proposed to characterize total charges of atoms that are exposed on the protein surface, which was shown to correlate with the viscosity of the protein.<sup>18</sup> SAP has been successfully applied to design a new sequence of an antibody that yielded a product that exhibited less aggregation.<sup>17</sup> However, care must be taken when mutating a protein sequence because APR sequences tend to be found in protein–protein binding sites, including complementary-determining regions (CDRs), and introducing mutation to the APR sequence may have adverse effect on therapeutic activity.<sup>19</sup>

Currently, computational tools that help formulation in screening for potential excipients are rare. Recent work by Cloutier et al.<sup>20</sup> has shown that preferential interaction coefficients ( $\Gamma_{23}$ ) calculated from MD simulation could be applied to select excipients that preferably interact with the protein (vs that remain in the bulk phase). Such excipients would be anticipated to prevent self-interaction of proteins and reduce viscosity. While it was shown that the  $\Gamma_{23}$  computed from MD simulation matches the experimental values, the correlation between  $\Gamma_{23}$  and the experimental viscosity was not clear. More recently, a machine learning model was developed for use in formulations with individual excipients based on  $\Gamma_{23}$  data in conjunction with a large number of mAb surface properties for five excipients and NaCl.<sup>21</sup> While the use of  $\Gamma_{23}$  in conjunction with machine learning opens up new avenues for formulation science, the model is based on MD simulations that require significant computational resources, especially when information on a large number of excipients or combination of excipients and buffers is required.

In this work, we describe the application of the SILCS (site-identification by ligand competitive saturation) approach<sup>22–24</sup> for ranking excipients for a given protein, here, monoclonal antibodies (mAb). Figure 1 shows the schematics of this



**Figure 1.** Overview of SILCS-Biologics workflow.

workflow, namely, SILCS-Biologics (Somani et al., submitted for publication). The workflow starts with the antibody 3D structure and generating SILCS FragMaps. The SILCS FragMaps encode the protein–functional group interaction pattern all across the protein surface while taking into account desolvation penalties and protein flexibility, allowing detailed description of protein–ligand interactions on any part of the

protein surface.<sup>23</sup> SILCS has been successfully applied for lead compound identification and optimization as well as the identification of novel ligand pockets in the structure-based small molecule drug development field.<sup>25–29</sup> The SILCS FragMaps are then used in antibody–antibody PPI docking and global docking of excipients on the antibody surface, utilizing the SILCS-PPI and SILCS-Hotspots methods, respectively.<sup>30,31</sup> Finally, the information gathered by the workflow is processed and compared with experimental data, such as viscosity, which allows future prediction of excipients that could modulate the properties of the antibody formulation.

The present context of excipient ranking builds upon this approach where the binding site is effectively the whole solvent exposed surface of the protein. Exploiting this along with the highly computationally efficient SILCS-MC approach allows SILCS to provide a detailed molecular level view of how well a given excipient “coats” the surface of a protein. This information may be used to understand how excipients may facilitate stabilization of the protein and, in conjunction with the SILCS-PPI approach, identify excipients that may bind to protein–protein self-interaction hot spots on the surface of the protein, thereby minimizing protein aggregation, a significant problem in protein-based drug formulation.

In this work, we apply the SILCS-Biologics approach to the NISTmAb<sup>32</sup> along with CNTO607 mAb,<sup>33</sup> focusing on the Fab in each case, to illustrate the utility of the workflow and demonstrate the types of information generated by the workflow. The NISTmAb is known to be highly stable such that it shows high concentratability and low viscosity at high concentrations. In contrast, the CNTO607 mAb aggregates at relatively low concentrations.<sup>34</sup> We show that PPI analysis indicates that the CDR of CNTO607 is more prone to be involved in self-association consistent with crystallographic and mutagenesis experiments.<sup>34</sup> In addition, with the NISTmAb, the number of excipients predicted to favorably interact with the antibody by SILCS-Biologics correlates with the decrease in viscosity, suggesting that the workflow can be used in screening a large number of excipients. Moreover, additional metrics from SILCS-Biologics are shown to correlate with the second virial coefficient of the Debye plot ( $B_{22}$ ), diffusion interaction parameter ( $k_D$ ), and the structure factor ( $S(0)_{\text{exp}}$ ) as well as with protein stability and allow for a molecular level interpretation of the events contributing to the impact of excipients on those properties.

## ■ MATERIALS AND METHODS<sup>A</sup>

**SILCS Simulations and FragMap Preparation.** Crystal structures of NISTmAb Fab, PDB: 5K8A and CNTO607 Fab, PDB: 3G6A, were used to initialize the SILCS simulations. SILCS FragMap simulations were performed using the SilcsBio software package version 2020.2 (SilcsBio, LLC) with the GROMACS simulation package.<sup>35</sup> The SILCS simulation protocol is previously described.<sup>28</sup> Briefly, the simulation systems involving protein, water, and eight solute molecules including benzene, propane, methanol, formamide, acetaldehyde, imidazole, methylammonium, and acetate were prepared. A total of 10 independent simulation systems were prepared with randomly positioned solute molecules at approximately 0.25 M each and with the sidechains of the solvent exposed protein residues systematically varied to improve convergence. The simulation systems were minimized and equilibrated for 1 ns with positional restraint on non-hydrogen protein atoms

using GROMACS simulation package.<sup>35</sup> Following 25 cycles of 200,000 steps of oscillating chemical potential grand canonical Monte Carlo (GCMC) sampling<sup>24,36</sup> of the water and solutes around the protein, the production simulations consist of 100 cycles of GCMC/MD simulation. During each cycle, 200,000 steps of GCMC simulation are performed, which drives the sampling of solute and water molecules, followed by 1 ns MD simulation of the entire system. Weak harmonic restraints with a force constant of 0.12 kcal/mol/Å were kept on all Cα atoms of the protein during the MD simulations. The unit cal stands for calorie, while 1 kcal equals to 1000 cal. The protein conformations and distributions of water and solutes were saved every 10 ps during the MD simulation phase for analysis. The protein, solutes, and water were described using the CHARMM36m protein force field,<sup>37,38</sup> the CHARMM general force field (CGenFF),<sup>39</sup> and the TIP3P water model modified for the CHARMM force field.<sup>40</sup>

The FragMaps were generated by binning the selected solute atoms into voxels of a 1 Å spaced grid spanning the simulation system. The voxel occupancy can be used to calculate grid-based free energy (GFE)<sup>41</sup> using the equation

$$\Delta G_i = -RT \ln(N_i/N_0)$$

where  $N_i$  is the observed voxel occupancy of the probe at grid point  $i$  and  $N_0$  is the expected voxel occupancy of the probe alone in the bulk for a given concentration. During the GCMC phase of the production simulation, the number of probe molecules fluctuates; therefore, the number of water molecules was used to correct the bulk concentration of the solute molecules.<sup>38</sup> GFE is a measure of free energy change for moving an atom from the bulk state to the grid point  $i$ . For example, if the GFE of a voxel near protein is −1.5 kcal/mol, then the probe atoms are about 12 times more likely to be found in that voxel than in the voxels that are far away from the protein (“bulk”) at room temperature.

**Protein–Protein Docking Using SILCS FragMap.** A global protein docking was performed using the SILCS FragMaps to determine the PPI preference (PPIP) using SilcsBio software package (SilcsBio, LLC). The details and the validation of SILCS-PPI protocol have been described previously.<sup>42</sup> Briefly, two sets of maps were prepared: SILCS FragMaps and the protein sidechain probability grid maps (PPGMaps) for both the receptor and the ligand proteins. The FragMaps and PPGMaps were computed from the same SILCS GCMC/MD trajectories described above. The PPGMaps were calculated by computing the voxel occupancy of selected sidechain atoms (see Table 1 of ref 42 for the selected sidechain atoms) and normalized by the maximum occupancy value, that is, from 0 to 1.

The SILCS-PPI protocol is designed to maximize the complementarity between the FragMaps from one protein and the PPGMaps from another protein. Both FragMaps and PPGMaps were preprocessed to reduce sampling noise. For the FragMaps, all voxels that are 5 Å away from the excluded area (i.e., protein interior) as defined by the SILCS exclusion map were removed since those grids are influenced by bulk phase behavior and only add noise to the sampling. For the ligand PPGMaps, all voxels that are overlapping with the exclusion map were removed since they occupy the repulsive core of the protein, and only PPGMaps that are accessible to the solvent or SILCS solute molecules are useful for PPI determination.

**Table 1.** Experimental Data for NISTmAb from the report by Xu et al.,<sup>46</sup> including Viscosity, Second Virial Coefficient of the Debye Plot ( $B_{22}$ ), Diffusion Interaction Parameter ( $k_D$ ), Structure Factor ( $S(0)_{\text{exp}}$ ), and Temperature at Which the Heat Capacity Starts to Change,  $\Delta T_{\text{onset}}$ , Versus Control

excipient	viscosity at 170 mg/mL (cp)	$B_{22}$ (mol mL/g <sup>2</sup> )	$k_D$ (mL/g)	$S(0)_{\text{exp}}$	$\Delta T_{\text{onset}}$ (°C)
arginine	29.06	$4.20 \times 10^{-5}$	−3.20	0.33	−0.02
trehalose	25.02	$3.24 \times 10^{-4}$	20.88	0.16	1.70
sucrose	24.43	$3.86 \times 10^{-4}$	19.84	0.11	3.08
glucose	21.21	$3.14 \times 10^{-4}$	23.49	0.09	3.47
mannitol	21.16	$2.95 \times 10^{-4}$	24.45	0.13	2.85
proline	18.69	$3.17 \times 10^{-4}$	22.10	0.10	1.01
glycine	16.32	$3.08 \times 10^{-4}$	26.20	0.07	1.90
alanine	15.82	$3.10 \times 10^{-4}$	21.20	0.09	3.10
control	12.2	$2.29 \times 10^{-4}$	19.80		0

The protein–protein docking was performed globally using a fast Fourier transform (FFT)-based algorithm.<sup>42</sup> The receptor FragMaps and PPGMaps were fixed in space, and the orientation of ligand FragMaps and PPGMaps were varied systematically over all possible orientations. For each orientation of the ligand protein, all possible translations of the ligand FragMaps and PPGMaps can be efficiently evaluated using FFT operations. The overlapped voxels from FragMaps and PPGMaps were multiplied and summed to yield the protein grid free energy (PGFE) score, as previously described.<sup>42</sup> Each docking calculation was performed with a 10° rotational angle interval on the 1 Å grid. The rotation angles were uniformly distributed to avoid the biases around the poles,<sup>43,44</sup> which resulted in 14,904 rotational angles. For each rotation angle, we stored the top 10 scoring solutions for further analysis, which resulted in 149,040 docked conformations. The details of the scoring function and FFT procedure can be found in ref 42.

PPIP map was calculated using a clustering analysis. Clustering analysis was performed on the total 149,040 docked poses. Due to the large number of poses, it is computationally untenable to use root-mean-square deviation based clustering algorithm. Instead, the center of mass (i.e.,  $x$ ,  $y$ , and  $z$  positions) and the orientation (i.e., Euler angles) of poses were used to compute the distance between the pose conformations. Two-step clustering was performed where the clusters are identified using the center of mass of the poses alone followed by another clustering analysis using the three Euler angles of poses that belong to the same cluster. The distance in angular space was measured by

$$d_{12} = \sqrt{\sum_{\theta}^{\alpha, \beta, \gamma} (\cos(\theta_1) - \cos(\theta_2))^2 + (\sin(\theta_1) - \sin(\theta_2))^2}$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  represent the three Euler angles. Cluster cutoffs of 10 Å and 0.5 (approximately 30°) were used for the center of mass and the Euler angles, respectively. In the case of the NISTmAb Fab, this procedure reduces the number of poses to 5508. After the clustering, the per-residue PPIP was computed as the number of contacts between the receptor and ligand atoms within a 5 Å cutoff and summed over all poses. The per-residue PPIP value is then normalized by the maximum per-residue PPIP value. This value is called the PPIP score, and higher values suggest that a residue is more likely involved in a PPI.



**Excipient Docking and Screening.** Excipient docking and screening is initiated with a docking algorithm, namely, SILCS-MC, which involves MC sampling of ligands in the field of the FragMaps for scoring.<sup>23</sup> Briefly, the SILCS-MC algorithm involves MC sampling of the ligand in translational, rotational, and torsional space in the field of FragMaps. The energy of a ligand conformation is evaluated by the combination of CGenFF intramolecular energies and ligand grid free energy (LGFE) score, which is the sum of atomic GFEs, as previously described.<sup>23,30</sup> The atomic GFE for each ligand atom is assigned by the FragMap voxel that the atom occupies. The protein structure is not explicitly included, but the exclusion map prevents the ligand from sampling the region where no solute or water molecules visited during the SILCS simulations, that is, the interior of the protein. This allows for rapid docking of the ligand while accounting for protein flexibility in a mean-field fashion as that information is embedded in the FragMaps and the exclusion map. LGFEs, which represent the predicted approximate free energy of binding of the excipients, have been shown to correlate well with the binding affinities of small, drug-like molecules to a range of proteins.<sup>30</sup>

To identify preferential binding of excipients on the protein surface, SILCS-Hotspots approach was used. In SILCS-Hotspots, the SILCS-MC method is extended to identify binding “hotspots” of a small molecule across the whole protein surface.<sup>31</sup> The detailed method and the validation of the method has been described previously.<sup>31</sup> Briefly, the entire FragMap space was divided into 14.14 Å<sup>3</sup> boxes. The box size is determined to be slightly larger than the SILCS-MC sampling region, which is a 10 Å radius sphere. In each box, the SILCS-MC sampling was performed for each excipient by randomly positioning it within the sphere of radius 10 Å centered at the middle of the sampling box. The excipient was then subjected to 10,000 MC steps at 300 K. During this phase of sampling, the molecular translations, orientation, and the torsion angle of an internal rotatable bond could change up to 1 Å, 180°, and 180°, respectively. This was followed by the 40,000 MC simulated annealing phase where the temperature of the system was gradually reduced from 300 to 0 K. During the annealing phase, the molecular translations, orientation, and the torsion angle of an internal rotatable bond could change up to 0.2 Å, 9°, and 9°, respectively. This process was repeated 1000 times for each excipient in each sampling box. The results from each individual sampling box were pooled together, and clustering analysis was performed.

To identify excipient binding “hotspots”, a center-of-mass based clustering analysis was performed with a simple clustering algorithm<sup>45</sup> where the conformer with the largest number of neighbors is identified as the cluster center, and its members were removed from the pool of conformations. We used adaptive clustering radius based on the molecular extent because different excipient molecules have different sizes. This process was repeated until no additional conformers remain. Once the cluster centroids for each excipient are identified, a second round of clustering was performed to identify the sites populated by different excipients. This was achieved by using the same clustering algorithm and a radius of 4 Å from which binding sites were identified that contain one or more members from the collection of excipients under study, for example, the binding sites where one or more types of excipient bind. The LGFE scores of each of the excipients in each site are then averaged with the sites ranked based on the

average LGFE scores. It should be noted that the site identification and the excipients included in each site are sensitive to the two clustering radii.

**Combining PPIP Maps and Protein–Excipient Docking.** SILCS-Biologics combines the PPIP and protein–excipient information, as previously described (Somani et al., submitted for publication). A total of 20 common excipient molecules ranging from amino acids to sugars were prepared and used in protein–excipient docking. The complete list of excipients are acetate, alanine, arginine, aspartic acid, citric acid, glucose, glutamic acid, glycine, histidine, lactate, lysine, malate, mannitol, phosphate, proline, sorbitol, succinate, sucrose, threonine, trehalose, and valine. The protonation states of the excipient molecules are considered at pH 6. The protonation state was chosen to match the experimental condition. Docking of each excipient is performed, as described in the previous section. Using the excipient binding site, a number of features associated with the antibody self PPIP and antibody–excipient interaction screening were extracted. For example, “total number of binding sites having LE < −0.25” is the number of binding poses per excipient having ligand efficiency (LE) ( $LE = LGFE / [\# \text{ of non-hydrogen atoms}]$ ) less than −0.25 kcal/mol, hence suggesting the number of binding poses that have relatively strong binding affinity.

**Experiments on NISTmAb.** Experiments and associated data on NISTmAb were previously presented,<sup>46</sup> with just a brief summary here. Experiments were performed at pH 6 in the presence of 25 mM histidine with excipient concentrations of 171 mM (Arg), 200 mM (Pro, Gly and Ala), 237 mM (mannitol), and 300 mM (sucrose, trehalose and glucose). The unit M stands for molar concentration of mol/L, while mM stands for concentration of 10<sup>−3</sup> mol/L. A series of NISTmAb solutions varying in mAb concentration were prepared for each excipient condition: the  $k_D$  and  $B_{22}$  results were measured from dilute mAb solutions (less than 20 mg/mL) and  $S(0)_{\text{exp}}$  values were measured from highest examined mAb concentrations (from 130 to 170 mg/mL), whereas the viscosity data was all obtained at 170 mg/mL. The viscosity of the histidine control was 12.2 cp. For more detailed list of the experimental conditions, see Table 1 of ref 46.

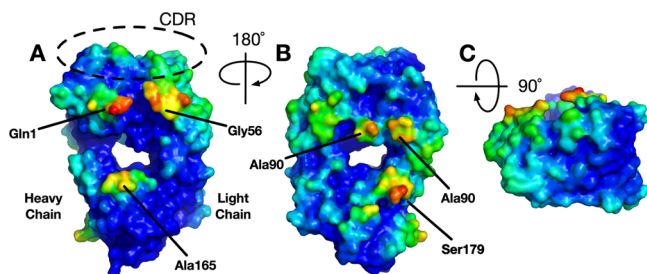
## RESULTS

### Antibody Fab Domain Self-Interaction PPIP Maps.

The SILCS-PPI protein–protein docking algorithm was used to examine the self-interaction preference of the Fab domain of the NISTmAb. While we have only used the Fab domain to reduce the computational cost, poses that would have been sterically inaccessible in full-length IgG1 antibody have been excluded. To filter those poses, the full-length IgG1 antibody structure (PDB: 1HZH) is overlaid on the receptor Fab pose (see Figure S1), and any poses in which the ligand Fab has a steric clash (having more than 2 ligands atoms within 5 Å of any full-length receptor atoms) with the second Fab in the IgG1 antibody are discarded. This is justified because the individual Fab and Fc domains are relatively rigid, while the relative orientation of Fab with respect to Fc as well as the second Fab varies significantly due to the flexible linkers. It is possible that the Fc domain may affect the FragMap densities by domain–domain interactions, particularly near the hinge region.<sup>32,47</sup> However, we expect such an effect will be limited to the Fab–Fc domain interface, and the FragMap densities near the elbow or CDR will not be affected. By focusing on the domain–domain interaction while not considering the domain

relative orientations, we can efficiently perform protein–protein docking and excipient hotspot mapping. In addition, the PPI and the excipient poses having steric clashes when overlaid with the full-length mAb were rejected, thereby reflecting some influences from the Fc domain on the SILCS metrics. This approach can be readily repeated over different full-length mAb conformations as more information about the orientation of Fab and Fc becomes available.

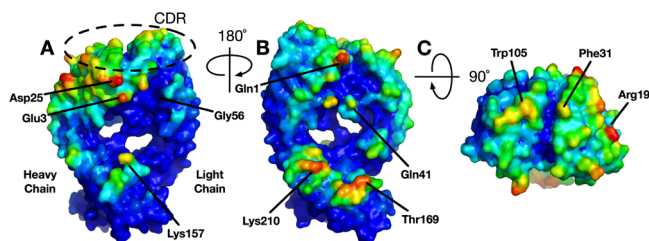
Figure 2 shows the self-interaction preference surface of the Fab domain of NISTmAb. The PPIP score is defined as the



**Figure 2.** NISTmAb Fab domain self-interaction preference surface. The interaction PPIP score is scaled to 0 (lower) to 1 (higher), and the color is assigned blue to red, respectively.

number of contacts per residue over the top 2000 docked poses sorted by the PGFE score. 5 Å was used as the cutoff distance for contact between any ligand and receptor non-hydrogen atoms. The number of contacts per residue is then normalized by the maximum number of contacts, resulting in an interaction preference score having values from 0 to 1. Interestingly, the NISTmAb does not show a strong interaction preference in the CDR region. The regions having the strongest PPIP are around the CDR on the sides of the Fab. This low interaction preference in the CDR is consistent with the experimental data showing this antibody to have low viscosity at high concentrations.<sup>46</sup> The PPIP self-interaction information can be used to introduce mutations to the antibody sequence to enhance the stability of the antibody.

We compare the protein self-interaction patterns of NISTmAb with CNTO607, an antibody which shows a high propensity of aggregation and with a well-characterized mechanism of aggregation.<sup>34,48</sup> CNTO607 mAb irreversibly precipitates at roughly 13 mg/mol, much lower than typical target concentration of therapeutic antibody formulations.<sup>48</sup> Figure 3 shows the self-interaction preference surface of the antibody CNTO607 Fab domain computed with the same procedure as the NISTmAb Fab domain. The CDR of CNTO607 shows much stronger preferences for self-

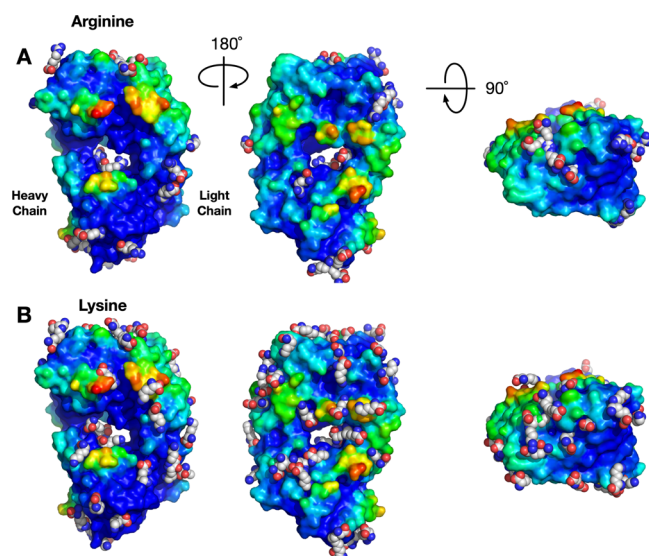


**Figure 3.** Antibody CNTO607 Fab domain self-interaction preference surface. The interaction PPIP score is scaled to 0 (lower) to 1 (higher), and the color is assigned blue to red, respectively.

interaction than the Fab domain of NISTmAb, where almost no residues in the CDR loops have significant self-interaction PPIP values. The SILCS-PPI simulations highlight the hydrophobic hotspot triad of  $_{103}\text{FWH}_{105}$  (same as F99, H100, and W100a in ref 48) in the heavy chain CDR3 which was identified to be responsible for aggregation based on X-ray crystallography and biophysical characterization of mutant proteins.<sup>34</sup> PPIP values for the F103, H104, and W105 were 0.91, 0.50, and 1.0, respectively, where 1.0 represents the maximum on the PPIP scale. Figure S2 shows the PPIP values of CNTO607 mapped on the solvent accessible surface and its amino acid sequence. In addition to the hydrophobic triad in the CDRs, the crystal structure of CNTO 607 Fab (PDB: 3G6A) showed salt bridge interactions between LC D49 and LC D50 (same as D50 and D51 of ref 46) in the Fv domain of one molecule with heavy chains K210 and K215 of another molecule (Figure S2). Mutations of the lysine residues disrupting the salt bridge interactions were found to improve solubility and viscosity,<sup>34</sup> indicating that these residues are also involved in self-association. The residues involved in the salt bridge are highlighted by the SILCS-PPI simulations (PPIP values of K210, K215, D49, and D50 were 0.25, 0.13, 0.41, and 0.28, respectively). It should be noted that the self-interaction PPIP value, in the current form, is a relative quantity, and the values from two different antibodies cannot be directly compared. Therefore, we cannot comment on whether one antibody is more prone to aggregation or not using just self-interaction preference. Nevertheless, if aggregation is observed, SILCS-PPI simulations can guide protein engineering efforts by identifying aggregation hotspot residues.

**Excipient–Fab Domain Interaction Analysis.** To investigate the excipient–protein interactions, a total of 20 commonly used excipients, such as amino acids and sugars, were selected and used in an excipient–Fab domain interaction analysis. For each excipient, global docking was performed, and the resulting poses were pooled together for clustering analysis, which resulted in the identification of multiple binding sites for each excipient. The cluster centroid positions of each excipient were clustered again to identify “hotspots” on the protein that are favored by one or more excipients. Similar analysis has been performed in protein–fragment interactions.<sup>31</sup> Figure 4 shows an example of hotspots of arginine (Arg) and lysine (Lys) given the energetic threshold of  $\text{LE} < -0.25$  kcal/mol. We chose LE for comparing energetics of different excipients because LE is independent of the size of the excipient molecules. Interestingly, Arg and Lys have quite similar chemical structures, but there are differences in the pattern of binding between the two excipients. We hypothesize that the difference is due to the bulkier guanidinium group on Arg, preventing Arg from fitting in a smaller pocket compared to Lys.

The information from the clustering approach with the excipients can be useful for understanding differential behavior of excipients or the competition between excipients and buffers for a given binding site. For example, the use of Lys or Arg as an excipient may lead to different outcomes during formulation, and it is not easy to rationalize the observation. In the NISTmAb Fab example, we found that all Arg-binding sites are also occupied by Lys given the energetic threshold, while there are additional sites to which only Lys binds. Figure S3 shows the binding sites for Lys that do not also bind Arg. Moreover, Lys binds to more regions of the protein predicted



**Figure 4.** Excipient–NISTmAb Fab domain interaction hotspots of Arg and Lys having LE  $< -0.25$  kcal/mol. The surface of NISTmAb Fab domain is colored using the Fab self-interaction preference score.

to be prone to self-interaction. From this observation, we hypothesized that properties based on two metrics, such as the total number of binding sites above a certain energetic threshold, could be used to correlate how well certain excipients “coat” the protein surface and impact aggregation and viscosity.

#### Correlation Analysis with Experimental Viscosity.

Beyond visual inspection, quantitative metrics of excipient–protein interactions can be examined. Table 2 shows a selected set of individual and combined metrics of various excipients from SILCS-Biologics. Only data for the subset of excipients that were included in the experimental measurements are presented here. See Table S1 with the complete list of excipients. The metrics include the number of binding sites to which each excipient binds based on either a  $-1$  kcal/mol LGFE or  $-0.25$  LE kcal/mol cutoff, the relative affinity analysis (RAA), which is the number of binding sites with a relative affinity within a factor of 1000 of the most favorable LGFE score for a given excipient, the number of binding sites with average PPIP score for residues within  $5 \text{ \AA}$  of the excipient greater than 0.1, a combination of the number of binding sites based on LE and PPIP, and the average LGFE and LE of the top 10 scoring poses of each excipient. Analysis of the values in Table 1 shows significant differences in all the properties between the excipients. For example, Arg binds to a

substantially lower number of sites based on LGFE, LE, and PPIP as compared to the other amino acids, with the number of sites similar to that of the some of the sugars. The predicted binding affinity shows sucrose and trehalose to be the most favorable, though taking into account the size of the molecules leads to them having the least favorable LE scores.

To understand the relationship of the SILCS-Biologics properties to the experimental observables, simple correlation analysis was performed on the metrics shown in Table 2 with the experimental data for the viscosity, the second virial coefficient of the Debye plot ( $B_{22}$ ), diffusion interaction parameter ( $k_D$ ), and the structure factor ( $S(0)_{\text{exp}}$ ) (Table 3). Notably, the correlations with the viscosity are quite high for many of the SILCS metrics. With metrics related to the number of binding sites occupied, there are systematically negative correlations, indicating that occupation of a larger number of binding sites is associated with lower viscosity. Reasonable correlations are also obtained with the LGFE and LE metrics averaged over the 10 most favorable binding poses. Interestingly, the viscosity is larger with more favorable LGFE values associated with the negative correlation, but the opposite occurs with LE with good correlations being obtained in both cases. Overall, the systematically strong correlations observed between viscosity and the various SILCS metrics indicate their potential utility for excipient selection.

The correlations between the SILCS metrics and  $B_{22}$ ,  $k_D$ , and  $S(0)_{\text{exp}}$  are systematically lower than those with the viscosity (Table 3).  $S(0)_{\text{exp}}$ , obtained from a small angle scattering experiment, has been shown to be of utility to study protein stability in dilute solutions as well as high concentrations.<sup>46</sup> This is consistent with the  $R^2$  value of  $S(0)_{\text{exp}}$  with viscosity,  $R^2 = 0.68$ , being higher than with either  $k_D$  and  $B_{22}$ . Lower correlations are obtained for  $B_{22}$  and  $k_D$  with viscosity,  $R^2 = 0.55$  and  $0.29$ , respectively. Although,  $B_{22}$  and  $k_D$  have been widely used in low-concentration experiments as an indicator of viscosity at higher concentrations typically used in formulations, the behavior of proteins in dilute solution and in concentrated formulation may be significantly different due to crowding effect, and predictions made in dilute solution may not be accurate.<sup>49–51</sup> The lower correlation of SILCS metrics with these biophysical measurements than the viscosity suggests the usefulness of SILCS metrics in predicting viscosity at high concentrations. The lower correlation with the biophysical measurements also appears to be in part due to the behavior of Arg, and that data in conjunction with the SILCS metrics may be used to obtain a molecular model leading to the changes in viscosity, as discussed below.

Shown in Figure 5 is the experimental viscosity versus “# of binding sites with LE  $< -0.25$ ” or “# of binding sites with PPIP

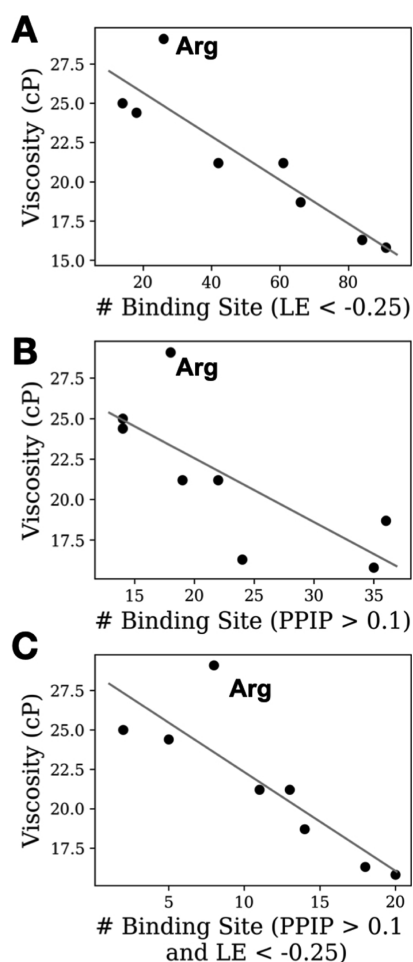
**Table 2.** Selected SILCS-Biologics Combined Properties Describing Protein–Excipient Interactions for the NISTmAb Fab Domain. LGFE and LE are in kcal/mol

excipient	# binding site (LGFE $< -1$ )	# binding site (LE $< -0.25$ )	# binding site (RAA $< 1000$ )	# binding site (PPIP $> 0.10$ )	# binding site (LE $< -0.25$ and PE $> 0.10$ )	Ave LGFE	Ave LE
arginine	62	26	13	18	8	−4.57	−0.38
trehalose	49	14	17	14	2	−7.35	−0.32
sucrose	50	18	14	14	5	−7.29	−0.32
glucose	94	61	51	22	13	−5.45	−0.46
mannitol	70	42	26	19	11	−5.92	−0.49
proline	131	66	86	36	14	−3.63	−0.46
glycine	105	84	105	24	18	−2.88	−0.58
alanine	137	91	137	35	20	−3.18	−0.53



**Table 3.** Correlation Analysis between Selected SILCS-Biologics Combined Properties Describing Protein–Excipient Interactions for the NISTmAb Fab Domain and Experimental Data from Ref 46 (Table 1)

		# binding site (LGFE < -1)	# binding site (LE < -0.25)	# binding site (RAA < 1000)	# binding site (PIIP > 0.10)	# binding site (LE < -0.25 and PIIP > 0.10)	Ave LGFE	Ave LE
viscosity	R	-0.83	-0.90	-0.90	-0.74	-0.85	-0.65	0.83
	R <sup>2</sup>	0.69	0.82	0.81	0.55	0.72	0.42	0.68
<i>B</i> <sub>22</sub>	R	0.16	0.18	0.24	0.10	0.07	-0.25	-0.06
	R <sup>2</sup>	0.02	0.03	0.06	0.01	0.00	0.06	0.00
<i>k</i> <sub>D</sub>	R	0.34	0.43	0.41	0.24	0.33	0.01	-0.41
	R <sup>2</sup>	0.11	0.18	0.17	0.06	0.11	0.00	0.17
<i>S</i> (0) <sub>exp</sub>	R	-0.50	-0.57	-0.57	-0.40	0.52	-0.17	0.47
	R <sup>2</sup>	0.25	0.33	0.33	0.16	0.27	0.03	0.22
without Arg		# binding site (LGFE < -1 kcal/mol)	# binding site (LE < -0.25)	# binding site (RAA < 1000)	# binding site (PE > 0.10)	# binding site (LE < -0.25 and PE > 0.10)	Ave LGFE	Ave LE
viscosity	R	-0.90	-0.98	-0.95	-0.82	-0.98	-0.98	0.92
	R <sup>2</sup>	0.81	0.96	0.91	0.68	0.95	0.96	0.84
<i>B</i> <sub>22</sub>	R	-0.49	-0.57	-0.44	-0.44	-0.55	-0.54	0.72
	R <sup>2</sup>	0.24	0.32	0.19	0.19	0.30	0.29	0.52
<i>k</i> <sub>D</sub>	R	0.25	0.51	0.28	0.11	-0.48	0.51	-0.77
	R <sup>2</sup>	0.06	0.26	0.08	0.01	0.23	0.26	0.59
<i>S</i> (0) <sub>exp</sub>	R	-0.70	-0.82	-0.73	-0.57	-0.83	-0.79	0.71
	R <sup>2</sup>	0.49	0.68	0.53	0.33	0.68	0.63	0.50

**Figure 5.** Correlation between the number of binding sites with (A) favorable binding affinity, (B) with high PPIIP, and (C) both favorable binding affinity and a high PPIIP value. The solid line is the best-fit curve using linear regression. The excipient Arg data points are indicated as they are an outlier, and the data for the plots are shown in Tables 1 and 2.

>0.1.” The strength of the correlations is obvious as is the presence of Arg as an outlier. With Arg as an excipient, the experimental viscosity is 29.1 cp significantly higher than that in the presence of the remaining excipients, though it should be noted that the viscosity in the presence of all the excipients is higher than that of the histidine (His) control value of 12.2 cp. To evaluate the impact of Arg on the predictability of the SILCS metrics, the correlation analysis was repeated in its absence, with the results presented in the lower portion of Table 3. The omission of Arg leads to improved correlations in all cases, with the improvements greatest with *B*<sub>22</sub> followed by *S*(0)<sub>exp</sub> and then *k*<sub>D</sub>.

Analysis of the SILCS metrics in Table 2 for the excipients allows for identification of potential indicators that may be used to flag this behavior a priori. With respect to the other experimentally studied amino acids, the differences with Arg are readily evident. The various numbers of binding sites occupied metrics are all significantly lower than the other amino acids, while the LE is less favorable. When all the amino acids are considered (Table S1), Arg still has the lowest number of binding sites based on LGFE and PPI metrics with Lys or His having the lowest values in the other SILCS metrics related to the number of binding sites. However, when the average over all 5 categories is taken, Arg yields a value of 25 versus 38 and 40 for Lys and His, respectively (Table S2). With respect to all the excipients, sucrose and trehalose typically have low values for the various number of binding site metrics. Notably, Arg, sucrose, and trehalose yield the highest viscosities observed in the experimental study (Table 1), indicating the potential predictive utility of this class of SILCS metrics. His alone yields the lowest viscosity, but this is only at 25 mM compared to the 171–300 mM concentrations of the other excipients in addition to His.

*T*<sub>onset</sub> the temperature at which significant changes in heat capacity are first observed, is an indicator of relative protein stability. In Table 4 the correlation between the SILCS-Biologics metrics and  $\Delta T_{\text{onset}}$  (changes in *T*<sub>onset</sub> values with the presence of various excipients as compared to the His control) is presented. In general, the correlations are low with the

**Table 4.** Correlation Analysis between Selected SILCS-Biologics Metrics Describing Protein–Excipient Interactions for the NISTmAb Fab Domain and Experimental  $\Delta T_{\text{onset}}$  from the Table 1 of ref 46

$\Delta T_{\text{onset}}$	# binding site (LGFE < −1 kcal/mol)	# binding site (LE < −0.25)	# binding site (RAA < 1000)	# binding site (PE > 0.10)	# binding site (LE < −0.25 and PE > 0.10)	Ave LGFE	Ave LE
<i>R</i>	0.09	0.24	0.18	−0.02	0.21	−0.24	−0.19
<i>R</i> <sup>2</sup>	0.01	0.06	0.03	0.00	0.04	0.06	0.04
Without Arg							
<i>R</i>	−0.18	0.00	−0.12	−0.26	0.08	−0.24	−0.02
<i>R</i> <sup>2</sup>	0.03	0.00	0.01	0.07	0.01	0.06	0.00

omission of Arg not leading to any systematic improvements. Analysis of the experimental data (Table 1) shows Arg to have a very small  $\Delta T_{\text{onset}}$  relative to the other studied amino acids. Thus, while the SILCS metrics are individually not correlative with  $\Delta T_{\text{onset}}$  for NISTmAb, in the context of the amino acids, the low number of binding sites for Arg suggests that these metrics may be representative of the impact of excipients on stability. Building on this, when multiple regression analysis is performed against the number of binding sites based on the LGFE < −1.0 kcal/mol and the average LGFE score, a *R*<sup>2</sup> value of 0.38 is obtained with respect to  $\Delta T_{\text{onset}}$ . The coefficients for the terms are 0.04 and −0.79, indicating that as the number of binding sites and the average LGFE becomes more favorable, the  $\Delta T_{\text{onset}}$  increases associated with increased stability of the protein. While the validity of performing multiple regression analysis with such a small number of data points is a concern, the result along with the physically intuitive model that increased interactions will lead to stabilization, do suggest the potential utility of the SILCS metrics for predicting the impact of excipients on protein stability. It should be noted that the  $\Delta T_{\text{onset}}$  will be affected by the domain that undergoes thermal melting, but the order of melting is ambiguous,<sup>46</sup> while the SILCS metrics are derived from the Fab domain–excipient interactions.

**Molecular Mechanism of Excipient Action.** To obtain insights into the molecular mechanism of excipient action, analysis of *B*<sub>22</sub> and *k*<sub>D</sub> was undertaken given the reasonable correlation of these terms in the absence of Arg. Both *B*<sub>22</sub> and *k*<sub>D</sub> terms are predictive of non-specific PPI, for example, positive *B*<sub>22</sub> and *k*<sub>D</sub> values are associated with reduced PPI, so they have been used to predict viscosity of protein formulation.<sup>52</sup> *B*<sub>22</sub> is measured via light scattering and, as such, includes information on direct PPI interactions while *k*<sub>D</sub>, being determined from the diffusion constant includes both thermodynamic (direct PPI) and hydrodynamic (frictional) contributions. Both *k*<sub>D</sub> and *B*<sub>22</sub> are typically measured in a dilute solution and the behavior of proteins in dilute solution versus in a concentrated formulation may be significantly different due to crowding effects, such that predictions made in dilute solution may not extrapolate to higher concentrations.<sup>49–51</sup> However, the differential information content in these experimental observables allows for them to be of utility to understand details of the impact of excipients on the viscosity in conjunction with the SILCS metrics.

*B*<sub>22</sub> typically shows a negative correlation with a various number of binding site terms and the average LGFE (Figures S4 and S5). This indicates that as an excipient binds to more sites on the Fab, there is a decrease in *B*<sub>22</sub> which would indicate an increase in attractive PPI interactions and, hence, increased viscosity. However, the variation of *B*<sub>22</sub> across the excipients is small (Table 1), indicating minimal impact of the excipients on direct PPI at the concentrations studied. Notably,

the outlier Arg has a *B*<sub>22</sub> value nearly an order of magnitude lower than that of the other excipients. This indicates that the low number of binding sites with Arg may contribute to a greater number of PPI. While sucrose and trehalose have similar trends for the number of binding sites, their more favorable LGFE scores suggest potentially longer resident times for these excipients that may offer additional inhibition of PPI leading to the larger *B*<sub>22</sub> and smaller viscosity values.

*k*<sub>D</sub> showed the largest correlation with the number of binding sites occupied based on LE (Figures S4 and S5) and combined LE and PPIP cutoffs with those correlations being lower than that with *B*<sub>22</sub>. Reasonable correlations were obtained with the average LGFE and LE terms. As with *B*<sub>22</sub>, the dependence of *k*<sub>D</sub> on the various binding site metrics was small though the relationship with average LGFE and LE was reversed with a more favorable LGFE leading to a lower *k*<sub>D</sub>. With *k*<sub>D</sub>, the lowest positive values besides the outlier Arg occur with sucrose and trehalose in contrast to *B*<sub>22</sub>, where these have the largest positive values. This suggests that these sugars are blocking direct PPI, with the lower *k*<sub>D</sub> values relative to the other excipients, excluding Arg, likely being due to hydrodynamics effects. This leads to a model where the more favorable binding affinity of sucrose and trehalose can protect against PPI though they are binding to a relatively small number of sites on the Fab, with these sugars in solution leading to higher viscosity via hydrodynamic effects. While generally it is thought that greater excipient–protein interactions lead to lowered viscosity, the present results suggest that the binding affinity of the excipient could also plays a role in reducing viscosity. With Arg, the small *B*<sub>22</sub> and negative *k*<sub>D</sub> suggest that the increased viscosity is due to the low number of binding sites that cannot be overcome by a more favorable binding affinity which would lead to the occurrence of additional PPI and increased viscosity. The negative *k*<sub>D</sub> (vs control = 19.8) for Arg may further indicate that hydrodynamic effects due to Arg in solution may further contribute to the increased viscosity. Interestingly, recent studies indicate the possible formation of Arg clusters in solution, leading to clusters containing 8–10 molecules.<sup>53</sup> Such phenomena may contribute the increased viscosity of Arg seen with the NISTmAb.

## DISCUSSION

Here, we apply SILCS-Biologics to both NISTmAb and CNT0607 mAb. The SILCS-Biologics workflow (Somani et al., submitted for publication) uses the SILCS technology to compute SILCS FragMaps with those FragMaps subsequently used in evaluation of the protein–protein self-interaction and protein–excipient interactions.<sup>31,42</sup> The protein–protein self-interaction patterns can provide insights on protein residues that are likely involved in self-interactions, which might be useful for designing a mutation that could lead to less



aggregation and improve viscosity. The protein–excipient interaction analysis is used to investigate the excipients that could potentially reduce aggregation and improve viscosity by “coating” the protein surface. The SILCS FragMap simulation is the most computationally expensive step in the SILCS-Biologics workflow, which requires 10 GPU nodes with at least 4 CPUs per node for 3–4 days. However, the subsequent PPI and protein–excipient interaction analysis steps do not require large compute resources, allowing screening of a large number of excipients.

The present results may be interpreted in the context of the preferential interaction parameter,  $\Gamma_{23}$ , a metric that indicates the excess number of excipient molecules interacting with the protein relative to being in solution.<sup>54,55</sup> A positive  $\Gamma_{23}$  indicates that the excipient has a higher local concentration on the protein surface, an effect interpreted to lead to increased protein solubility as the potential PPI contact area is decreased due to the presence of the excipients on the protein surface. In the opposite case, with a negative  $\Gamma_{23}$ , the concentration of the excipient at the protein surface is low relative to in solution, such that the protein is more favorably hydrated, leading to decreased protein solubility as more PPI interactions can occur. This behavior was observed with Arg and the relatively insoluble proteins human MAGOH, fragment 1–153 of murine Ref2-1 and WW domains 3 and 4 from *Drosophila* Su(dx).<sup>56</sup> However, subsequent studies on lysozyme and Arg indicate the reverse outcome where  $\Gamma_{23}$  was decreased.<sup>57</sup> These examples highlight the challenge of protein formulation where observations and models developed for a given set of protein and conditions do not translate to different proteins. We hypothesize that this is due to local effects occurring on the protein surface, where the occupation of specific regions on the protein surface by particular excipients impacts the potential occurrence of PPI. Thus, it is essential to obtain a descriptor analogous to  $\Gamma_{23}$  that varies across the protein surface. The SILCS-Biologics methodology directly addresses this by allowing for the full distribution of excipient binding sites on the protein surface to be determined including the predicted binding affinities of the excipients for each site.

Notable is the high correlation of the SILCS-Biologics with the published experimental data. A general trend of an increased number of binding sites on the protein by an excipient leading to lower viscosity is observed indicating that increased coating of the protein surface does lead to a lower viscosity. However, based on the anomalous behavior of Arg along with the results in sucrose and trehalose, the balance between the number of predicted binding sites and the predicted binding affinity, based on the average LGFE, further indicates a model where the lowered number of binding sites may be partially overcome by a more favorable binding affinity, minimizing an increase in viscosity. Finally, while the individual SILCS metrics did not show correlation with the experimental data on protein stability, the combination of the predicted total number of binding sites with LGFE <−1 kcal/mol combined with the predicted binding affinity based on the average LGFE does yield a reasonable correlation. Overall, these results indicate the potential utility of SILCS-Biologics for the rational prediction of excipients to both maintain lower viscosity and improve protein stability during the formulation process.

The current SILCS-Biologics workflow has a few limitations. First, the SILCS simulations are performed with a fixed charge model where the charges are determined with a fixed pH

condition. Given experimental formulations evaluate protein–excipients in a wide range of pH conditions, SILCS FragMaps computed in a single protonation state is limiting. Approaches that can overcome the limitation of single protonation state are being explored. Second, the self-interaction PPI score is expressed in relative terms, which makes it difficult to compare the magnitude of self-interaction preference from one protein to another. Finally, it needs to be emphasized that given the variability of the physical properties of antibodies, the ability to extrapolate the present results from the NISTmAb to other mAb is likely limited. Further studies on a wider range of mAb on which extensive experimental evaluation has been undertaken is required. The data used in this work is available upon request.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.0c00775>.

Example of protein–protein docking pose filtering, CNTO607 surface and sequences colored by PPIP values, lysine exclusive binding sites on the NIST Fab domain, correlation plots of SILCS metrics and experimental data, and calculated and average SILCS metrics for the NIST Fab domain (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

Alexander D. MacKerell, Jr. — Computer Aided Drug Design Center, Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201, United States; [orcid.org/0000-0001-8287-6804](https://orcid.org/0000-0001-8287-6804); Email: [alex@outerbanks.umaryland.edu](mailto:alex@outerbanks.umaryland.edu)

### Authors

Sunhwan Jo — SilcsBio, LLC, Baltimore, Maryland 21202, United States; [orcid.org/0000-0002-4104-6473](https://orcid.org/0000-0002-4104-6473)

Amy Xu — NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

Joseph E. Curtis — NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

Sandeep Somani — Discovery Sciences, Janssen Research and Development (Janssen R&D), Spring House, Pennsylvania 19477, United States

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.0c00775>

### Notes

The authors declare the following competing financial interest(s): ADM Jr is co-founder and Chief Scientific Officer of SilcsBio, LLC. SJ is an employee of SilcsBio LLC when the studies were performed.

## ■ ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant R43GM130198. The authors acknowledge computer time and resources from the Computer-Aided Drug Design (CADD) Center at the University of Maryland, Baltimore.

## ■ ADDITIONAL NOTE

<sup>a</sup>Certain commercial equipment, instruments, or materials (or suppliers, or software, ...) are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

## ■ REFERENCES

- (1) Ecker, D. M.; Jones, S. D.; Levine, H. L. The therapeutic monoclonal antibody market. *mAbs* **2015**, *7*, 9–14.
- (2) Kintzing, J. R.; Filsinger Interrante, M. V.; Cochran, J. R. Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. *Trends Pharmacol. Sci.* **2016**, *37*, 993–1008.
- (3) Walsh, G. Biopharmaceutical benchmarks 2010. *Nat. Biotechnol.* **2010**, *28*, 917–924.
- (4) Mullard, A. 2016 FDA drug approvals. *Nat. Rev. Drug Discovery* **2017**, *16*, 73–76.
- (5) Mueller, C.; Altenburger, U.; Mohl, S. Challenges for the pharmaceutical technical development of protein coformulations. *J. Pharm. Pharmacol.* **2017**, *70*, 666–674.
- (6) Cauchy, M.; D'aoust, S.; Dawson, B.; Rode, H.; Hefford, M. A. Thermal Stability: A Means to Assure Tertiary Structure in Therapeutic Proteins. *Biologicals* **2002**, *30*, 175–185.
- (7) Huus, K.; Havelund, S.; Olsen, H. B.; van de Weert, M.; Frokjaer, S. Thermal dissociation and unfolding of insulin. *Biochemistry* **2005**, *44*, 11171–11177.
- (8) Shire, S. J. Formulation and manufacturability of biologics. *Curr. Opin. Biotechnol.* **2009**, *20*, 708–714.
- (9) 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.
- (10) Roberts, C. J. Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol.* **2014**, *32*, 372–380.
- (11) Roberts, C. J.; Das, T. K.; Sahin, E. Predicting solution aggregation rates for therapeutic proteins: approaches and challenges. *Int. J. Pharm.* **2011**, *418*, 318–333.
- (12) Agrawal, N. J.; Kumar, S.; Wang, X.; Helk, B.; Singh, S. K.; Trout, B. L. Aggregation in Protein-Based Biotherapeutics: Computational Studies and Tools to Identify Aggregation-Prone Regions. *J. Pharm. Sci.* **2011**, *100*, 5081–5095.
- (13) Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **2004**, *22*, 1302–1306.
- (14) Tartaglia, G. G.; Cavalli, A.; Pellarin, R.; Cafisch, A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* **2005**, *14*, 2723–2734.
- (15) Tartaglia, G. G.; Pawar, A. P.; Campioni, S.; Dobson, C. M.; Chiti, F.; Vendruscolo, M. Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **2008**, *380*, 425–436.
- (16) Tartaglia, G. G.; Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **2008**, *37*, 1395–1401.
- (17) 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.
- (18) 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.
- (19) Chennamsetty, N.; Voynov, V.; Kayser, V.; Helk, B.; Trout, B. L. Prediction of Aggregation Prone Regions of Therapeutic Proteins. *J. Phys. Chem. B* **2010**, *114*, 6614–6624.
- (20) 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. Pharm.* **2019**, *16*, 3657–3664.
- (21) Cloutier, T. K.; Sudrik, C.; Mody, N.; Sathish, H. A.; Trout, B. L. Machine Learning Models of Antibody-Excipient Preferential Interactions for Use in Computational Formulation Design. *Mol. Pharm.* **2020**, *17*, 3589.
- (22) Guvench, O.; MacKerell, A. D., Jr. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435–12.
- (23) Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell, A. D., Jr. Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach. *J. Chem. Inf. Model.* **2013**, *53*, 3384–3398.
- (24) Lakkaraju, S. K.; Raman, E. P.; Yu, W.; MacKerell, A. D., Jr. Sampling of Organic Solutes in Aqueous and Heterogeneous Environments Using Oscillating Excess Chemical Potentials in Grand Canonical-like Monte Carlo-Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2281–2290.
- (25) Cao, X.; Yap, J. L.; Newell-Rogers, M. K.; Peddaboina, C.; Jiang, W.; Papaconstantinou, H. T.; Jupiter, D.; Rai, A.; Jung, K.-Y.; Tubin, R. P.; Yu, W.; Vanommelaeghe, K.; Wilder, P. T.; MacKerell, A. D., Jr.; Fletcher, S.; Smythe, R. W. The novel BH3  $\alpha$ -helix mimetic JY-1-106 induces apoptosis in a subset of cancer cells (lung cancer, colon cancer and mesothelioma) by disrupting Bcl-xL and Mcl-1 protein-protein interactions with Bak. *Mol. Cancer* **2013**, *12*, 42.
- (26) Cardenas, M. G.; Yu, W.; Beguelin, W.; Teater, M. R.; Geng, H.; Goldstein, R. L.; Oswald, E.; Hatz, K.; Yang, S.-N.; Cohen, J.; Shakhovich, R.; Vanommelaeghe, K.; Cheng, H.; Liang, D.; Cho, H. J.; Abbott, J.; Tam, W.; Du, W.; Leonard, J. P.; Elemento, O.; Cerchietti, L.; Cierpicki, T.; Xue, F.; MacKerell, A. D., Jr.; Melnick, A. M. Rationally designed BCL6 inhibitors target activated B cell diffuse large B cell lymphoma. *J. Clin. Invest.* **2016**, *126*, 3351–3362.
- (27) Cheng, H.; Linhares, B. M.; Yu, W.; Cardenas, M. G.; Ai, Y.; Jiang, W.; Winkler, A.; Cohen, S.; Melnick, A.; MacKerell, A. D., Jr.; Cierpicki, T.; Xue, F. Thiourea-Based Inhibitors of the B-Cell Lymphoma 6 (BCL6) BTB Domain via NMR- Based Fragment Screening and Computer-Aided Drug Design. *J. Med. Chem.* **2018**, *61*, 7573–7588.
- (28) Raman, E. P.; Lakkaraju, S. K.; Denny, R. A.; MacKerell, A. D., Jr. Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by Ligand competitive saturation approaches. *J. Comput. Chem.* **2017**, *38*, 1238–1251.
- (29) Yu, W.; Lakkaraju, S. K.; Raman, E. P.; Fang, L.; MacKerell, A. D., Jr. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. *J. Chem. Inf. Model.* **2015**, *55*, 407–420.
- (30) Ustach, V. D.; Lakkaraju, S. K.; Jo, S.; Yu, W.; Jiang, W.; MacKerell, A. D., Jr. Optimization and Evaluation of Site-Identification by Ligand Competitive Saturation (SILCS) as a Tool for Target-Based Ligand Optimization. *J. Chem. Inf. Model.* **2019**, *59*, 3018–3035.
- (31) MacKerell, A. D., Jr.; Jo, S.; Lakkaraju, S. K.; Lind, C.; Yu, W. Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (SILCS-Hotspots). *Biochim. Biophys. Acta, Gen. Subj.* **2020**, *1864*, 129519.
- (32) Castellanos, M. M.; Howell, S. C.; Gallagher, D.; Curtis, J. E. Characterization of the NISTmAb Reference Material using Small-Angle Scattering and Molecular Simulation. *Anal. Bioanal. Chem.* **2018**, *410*, 2141–2159.
- (33) Teplyakov, A.; Obmolova, G.; Wu, S.-J.; Luo, J.; Kang, J.; O'Neil, K.; Gilliland, G. L. Epitope mapping of anti-interleukin-13 neutralizing antibody CNTO607. *J. Mol. Biol.* **2009**, *389*, 115–123.
- (34) Bethea, D.; Wu, S.-J.; Luo, J.; Hyun, L.; Lacy, E. R.; Teplyakov, A.; Jacobs, S. A.; O'Neil, K. T.; Gilliland, G. L.; Feng, Y. Mechanisms

of self-association of a human monoclonal antibody CNT0607. *Protein Eng., Des. Sel.* **2012**, *25*, 531–538.

(35) 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 super-computers. *SoftwareX* **2015**, *1–2*, 19–25.

(36) Lakkaraju, S. K.; Yu, W.; Raman, E. P.; Hershfeld, A. V.; Fang, L.; Deshpande, D. A.; MacKerell, A. D., Jr. Mapping functional group free energy patterns at protein occluded sites: nuclear receptors and G-protein coupled receptors. *J. Chem. Inf. Model.* **2015**, *55*, 700–708.

(37) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone  $\phi$ ,  $\psi$  and side-chain  $\chi(1)$  and  $\chi(2)$  dihedral angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(38) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Iler, H. G. u.; MacKerell, A. D., Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(39) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(40) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. Solvent-Induced Forces between Two Hydrophilic Groups. *J. Phys. Chem.* **1994**, *98*, 2198–2202.

(41) Raman, E. P.; Yu, W.; Guvench, O.; MacKerell, A. D., Jr. Reproducing crystal binding modes of ligand functional groups using Site-Identification by Ligand Competitive Saturation (SILCS) simulations. *J. Chem. Inf. Model.* **2011**, *51*, 877–896.

(42) Yu, W.; Jo, S.; Lakkaraju, S. K.; Weber, D. J.; MacKerell, A. D., Jr. Exploring protein-protein interactions using the Site-Identification by Ligand Competitive Saturation (SILCS) methodology. *Proteins* **2019**, *87*, 289–301.

(43) Lovisolo, L.; da Silva, E. A. B. Uniform distribution of points on a hyper-sphere with applications to vector bit-plane encoding. *IEE Proc. Vis. Image Signal Process.* **2001**, *148*, 187–193.

(44) Solernou, A.; Fernandez-Recio, J. Protein docking by Rotation-Based Uniform Sampling (RotBUS) with fast computing of intermolecular contact distance and residue desolvation. *BMC Bioinf.* **2010**, *11*, 352.

(45) Zhang, Y.; Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* **2004**, *25*, 865–871.

(46) Xu, A. Y.; Castellanos, M. M.; Mattison, K.; Krueger, S.; Curtis, J. E. Studying Excipient Modulated Physical Stability and Viscosity of Monoclonal Antibody Formulations Using Small-Angle Scattering. *Mol. Pharm.* **2019**, *16*, 4319–4338.

(47) Koenig, P.; Lee, C. V.; Walters, B. T.; Janakiraman, V.; Stinson, J.; Patapoff, T. W.; Fuh, G. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E486–E495.

(48) Wu, S.-J.; Luo, J.; O'Neil, K. T.; Kang, J.; Lacy, E. R.; Canziani, G.; Baker, A.; Huang, M.; Tang, Q. M.; Raju, T. S.; Jacobs, S. A.; Teplyakov, A.; Gilliland, G. L.; Feng, Y. Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Eng., Des. Sel.* **2010**, *23*, 643–651.

(49) Sorret, L. L.; DeWinter, M. A.; Schwartz, D. K.; Randolph, T. W. Challenges in Predicting Protein-Protein Interactions from Measurements of Molecular Diffusivity. *Biophys. J.* **2016**, *111*, 1831–1842.

(50) 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.

(51) Woldeyes, M. A.; Qi, W.; Razinkov, V. I.; Furst, E. M.; Roberts, C. J. How Well Do Low- and High-Concentration Protein

Interactions Predict Solution Viscosities of Monoclonal Antibodies? *J. Pharm. Sci.* **2019**, *108*, 142–154.

(52) James, S.; McManus, J. J. Thermal and Solution Stability of Lysozyme in the Presence of Sucrose, Glucose, and Trehalose. *J. Phys. Chem. B* **2012**, *116*, 10182–10188.

(53) Das, U.; Hariprasad, G.; Ethayathulla, A. S.; Manral, P.; Das, T. K.; Pasha, S.; Mann, A.; Ganguli, M.; Verma, A. K.; Bhat, R.; Chandrayan, S. K.; Ahmed, S.; Sharma, S.; Kaur, P.; Singh, T. P.; Srinivasan, A. Inhibition of protein aggregation: supramolecular assemblies of arginine hold the key. *PLoS One* **2007**, *2*, e1176.

(54) Arakawa, T.; Timasheff, S. N. Theory of protein solubility. *Methods Enzymol.* **1985**, *114*, 49–77.

(55) Arakawa, T.; Bhat, R.; Timasheff, S. N. Preferential interactions determine protein solubility in three-component solutions: the MgCl<sub>2</sub> system. *Biochemistry* **1990**, *29*, 1914–1923.

(56) Golovanov, A. P.; Hautbergue, G. M.; Wilson, S. A.; Lian, L.-Y. A simple method for improving protein solubility and long-term stability. *J. Am. Chem. Soc.* **2004**, *126*, 8933–8939.

(57) Schneider, C. P.; Trout, B. L. Investigation of cosolute-protein preferential interaction coefficients: new insight into the mechanism by which arginine inhibits aggregation. *J. Phys. Chem. B* **2009**, *113*, 2050–2058.