Anisotropic coarse-grain Monte Carlo simulations of lysozyme, lactoferrin, and NISTmAb by precomputing atomistic models

Harold W Hatch, Christina Bergonzo, Marco A Blanco, Guangcui Yuan, Sergei Grudinin, Mikael Lund, Joseph E Curtis, Alexander V Grishaev, Yun Liu, Vincent K Shen

Anisotropic coarse-grain Monte Carlo simulations of lysozyme, lactoferrin, and NISTmAb by precomputing atomistic models

Harold W. Hatch,[1, a)] Christina Bergonzo,[2, 3] Marco A. Blanco,[4] Guangcui Yuan,[5] Sergei Grudinin,[6] Mikael Lund,[7] Joseph E. Curtis,[8] Alexander V. Grishaev,[2, 3] Yun Liu,[5, 9] and Vincent K. Shen[1]

[1)] *Chemical Informatics Research Group, Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8380, USA*

[2)] *Institute for Bioscience and Biotechnology Research, Rockville, Maryland, USA*

[3)] *Biomolecular Structure and Function Group, Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8380, USA*

[4)] *Discovery Pharmaceutical Sciences, Merck Research Laboratories, Merck & Co., Inc, West Point, Pennsylvania 19486, USA*

[5)] *Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*

[6)] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*

[7)] *Division of Computational Chemistry, Lund University, Lund, Sweden*

[8)] *NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*

[9)] *Center for Neutron Science, Department of Chemical and Biomolecular Engineering, College of Engineering, University of Delaware, Newark, Delaware 19711, USA*

(Dated: 5 September 2024)

We develop a multiscale coarse-grain model of the NIST Monoclonal Antibody Reference Material 8671 (NISTmAb) to enable systematic computational investigations of high-concentration physical instabilities such as phase separation, clustering, and aggregation. Our multiscale coarse-graining strategy captures atomic-resolution interactions with a computational approach that is orders of magnitude more efficient than atomistic models, assuming the biomolecule can be decomposed into one or more rigid bodies with known, fixed structures. This method reduces interactions between tens of thousands of atoms to a single anisotropic interaction site. The anisotropic interaction between unique pairs of rigid bodies is precomputed over a discrete set of relative orientations and stored, allowing interactions between arbitrarily oriented rigid bodies to be interpolated from the precomputed table during coarse-grained Monte Carlo simulations. We present this approach for lysozyme and lactoferrin as a single rigid body and for the NISTmAb as three rigid bodies bound by a flexible hinge with an implicit solvent model. This coarse-graining strategy predicts experimentally measured radius of gyration and second osmotic virial coefficient data, enabling routine Monte Carlo simulation of medically relevant concentrations of interacting proteins while retaining atomistic detail. All methodologies used in this work are available in the open-source software Free Energy and Advanced Sampling Simulation Toolkit (FEASST). (This is an author reprint of https://doi.org/10.1063/5.0224809).

I. INTRODUCTION

As annual pharmaceutical sales of antibodies total nearly 10^{11} United States dollars[1,2] to treat cancer, arthritis, psoriasis, asthma, and other conditions,[3] improvements in processing, storage, or delivery to patients could have a large economic and quality of life impact.[4] While antibodies are commonly delivered to patients by intravenous administration,[5] other desirable outpatient and convenient delivery options include subcutaneous and subdermal injections. Such alternative routes of administration require high protein concentration formulations to meet the required dosage.[6] These formulations may encounter physical instabilities such as phase separation, high viscosity, and aggregation.[7–9] Models of protein-protein interactions[10] may be used to screen pharmaceutical candidates to avoid these instabilities.

The focus of this work is developing a computational method to simulate monoclonal antibodies (mAbs) in high concentration formulations.[11] We use the NIST Monoclonal Antibody Reference Material 8671 (NISTmAb)[12] as a model protein to demonstrate our method to better understand and avoid these high-concentration physical instabilities. NISTmAb is a non-proprietary mAb available to the public for fundamental scientific study.[13] We also model lysozyme and lactoferrin for further validation of the methodology and comparison to experiments with increasing size and complexity.

All-atom (AA) modeling of mAb-mAb interactions is computationally costly and largely prohibitive due to the large size of the solvated system.[14,15] In previous studies, AA explicit solvent molecular dynamics (MD) simulations of 1.2×10^7 atoms were performed for hundreds of nanoseconds.[16] The special purpose Anton 2 supercomputer[17] was also used to simulate an AA model of *E. coli* cytoplasm up to microsecond timescales.[18] In this work, we hope to develop a multiscale coarse-grained

[a)] Electronic mail: harold.hatch@nist.gov

(CG) modeling strategy that can be deployed by the pharmaceutical industry to quickly model candidates in early stage development without the use of specialized computers. We use AA models of proteins with thousands of interaction sites to develop CG models[19] of proteins with only a few interaction sites[20–25] and use Monte Carlo (MC) simulations to validate those models against experimental data.

There are numerous examples of CG computational models of mAbs described via sets of discrete interaction sites,[26] typically ranging in number between 3 and 12.[10,27–44] These CG models can be used to compute and predict second osmotic virial coefficients (B_2),[45,46] osmotic compressibility,[47] viscosity,[48] self-diffusion,[49] x-ray scattering,[50] and structure factors.[51] While many of these models are rigid, others include flexibility in the hinge region.[52] In this work, AA MD simulations with explicit solvent are used to parameterize the flexible inter-domain linkers in the NISTmAb CG model. In addition, while nearly all previous mAb models used isotropic interactions, we use anisotropic interactions in this work. While those require additional computational cost and complexity, our model enables sub-nm resolution despite having only three anisotropic interaction sites, one for each of the antibody domains.

In a previous CG approach, Brownian dynamics modeled proteins as rigid bodies in a continuum solvent, where the potential was precomputed on cubic grids.[53,54] Precomputation introduces an additional computationally expensive step prior to the multi-molecule simulations while accelerating these subsequent many-protein simulations by several orders of magnitude. Precomputed potentials may be considered CG models with the CG resolution determined by the grid spacing. This approach was also used to account for molecular flexibility[55,56] and macromolecular crowding.[57] Instead of precomputing potentials on a cubic grid in this work, we precompute pair-wise interactions.

Precomputation of pair-wise interactions as a function of their relative orientation has been used previously to simulate triangles[58] and squares[59] with rounded edges, cylinders,[60,61] and supertoroids.[62] In addition, docking-based fast Fourier transform (FFT) approaches may enumerate low-energy relative orientations of rigid proteins for long timescale MC simulations that transition between these states.[63] In this work, we precompute pair-wise interactions of arbitrary 3-dimensional rigid objects and apply this method to simulate proteins. This CG strategy relies upon an existing detailed model for a complex biomolecule and introduces an additional approximation and assumption. The approximation is that the biomolecule can be reasonably decomposed into one or more rigid bodies with known, fixed structure(s) and that the interaction energy between these rigid bodies can be computed for a fixed relative separation distance and relative orientation without the need for ensemble averages (e.g., implicit solvent model).

For precomputing pair-wise interactions, there is flexibility in the choice of the AA potential. Many AA potentials perform well when a single protein at high dilution is compared against experimental data with explicit solvent. However, to reduce the computational cost of precalculation, implicit solvent models allow the interaction energy of a fixed relative orientation to be computed quickly without the need for statistical-mechanical ensemble averages with explicit solvent. In this work, we utilize an implicit solvent model inspired by the protein docking community and parameterized from experimental B_2 data.[64]

For various reasons, experimental measurement of B_2 values was used to validate the CG models. B_2 is the orientationally-averaged interaction between two particles[65] and is one way to compare the relative average attraction or repulsion in different models. B_2 values are often used to characterize protein phase behavior.[45,66,67] A more positive B_2 indicates more repulsions, while a more negative B_2 indicates more attractions. In addition, experimental measurement of B_2 from light and small-angle scattering intensity allows comparison with theoretical models[46] and prediction of liquid-liquid phase separation.[68] Furthermore, extended corresponding states[69] utilize B_2 to collapse phase diagrams of short-range models. In the context of comparing tabular anisotropic potentials at various resolutions, B_2 was also shown to be a useful metric for cubic[59] and cylindrical[60,61] colloids. Fitting B_2 of models to experimental data is also simplified by the prediction of B_2 using derivatives.[70–72] On the other hand, the prediction of high-concentration behavior with low-concentration metrics, such as B_2, can be difficult. This may be due to some interactions that may not contribute greatly to B_2 at low concentrations,but may play an important role at higher concentrations.[38,49,73,74]

MC simulations play a central role in this article for two reasons. The first reason is that precomputation of the AA potential is simplified when only the potential energy is required, as opposed to the forces and torques required in MD. The second reason is that Mayer-sampling MC (MSMC)[75] is an efficient algorithm for computing B_2 and validating the CG models.

This article applies a CG methodology for precomputing interactions between rigid bodies to lysozyme and lactoferrin as a single rigid body in Section II. Varying levels of precomputed angular and distance resolutions are compared against B_2 using MSMC in Sections III. The NISTmAb CG model of three rigid bodies connected by a flexible hinge region is then simulated using MSMC in Section IV. Finally, conclusions and future work are discussed in Section V.

II. LYSOZYME AND LACTOFERRIN AS A SINGLE RIGID BODY

This section outlines the CG methodology by representing lysozyme and lactoferrin as isolated rigid bodies.

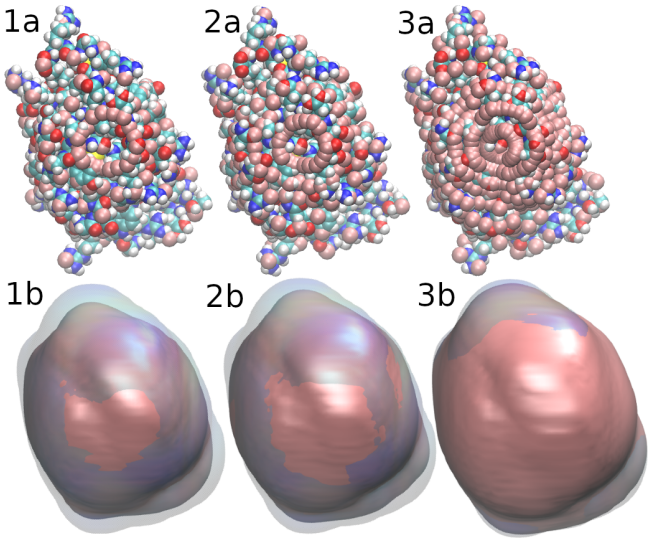

FIG. 1: (1a)-(3a) Lysozyme (PDB ID 4LYT) structures are shown with default VMD[76] atom name coloring. In addition, 3 Å mauve-colored beads were embedded in the surface with a solid-angular resolution of (1) 18° with $n_k = 10$, (2) 12° with $n_k = 15$, and (3) 10° with $n_k = 18$. (1b)-(3b) The blue transparent surface shows the VMD surface representation generated by rolling a 3 Å bead about all atoms. The solid mauve surface shows the surface generated from the mauve beads for each resolution above.

The major assumption we make with this CG approach is that the pair interaction between CG sites is effectively described with six rigid body degrees of freedom. Although in this article, we assume the atomic structure of the protein is rigid, the potential function may in principle contain some integrated flexible degrees of freedom or multiple configurations.[55] By assuming a rigid atomic structure, the AA interactions between two rigid bodies can be computed and stored as a function of their relative orientation. This stored table of interactions is referred to in this work as the CG model, where CG simulations interpolate from the stored AA interactions during MC simulations, and each rigid body is represented by a single anisotropic site. Because the computational cost of one anisotropic site roughly corresponds to that of three isotropic sites, a CG model of a protein with 6×10^3 atoms is over 10^3 times faster than an AA simulation. The AA interactions may be precomputed over a finite number of relative orientations. This is referred to as the resolution of the CG model.

A. Orientational resolution and uniqueness

We will begin by illustrating varying levels of orientational resolution of CG models. The top portion of Fig. 1 shows the rigid structure of lysozyme from the Protein Data Bank (PDB) ID 4LYT[77] with mauve beads embedded in the surface to visualize orientations for the anisotropic CG model at various levels of angular resolution. Each mauve bead represents a single polar and azimuthal angle emanating from the center of mass of the lysozyme onto the surface. In Fig. 1, the z-axis and the pole of the azimuthal angle are perpendicular to the page surface, as evidenced by the mauve beads' approximately circular surface ring structures.

The orientational or angular resolution is determined by the parameter n_k, the number of orientations per 180° in each solid angle. In this work, the n_k orientations were linearly spaced in each angle to enable fast linear interpolation between them. As n_k increases, the number of mauve beads increases in Fig. 1. Therefore, a larger n_k better represents the AA structure at the computational cost of a larger number of relative orientations. For $n_k = 12$, there are $180°/12 = 15°$ between each angle, corresponding to an approximate 5 Å arc length of a 19 Å radius sphere.[78]

The lower half of Fig. 1 shows the surface of the atomic structure as seen by a 3 Å bead with a translucent blue surface, while the CG model is shown in mauve. For $n_k = 10$ (18 degree resolution) on the left, the AA surface is significantly larger than the CG surface. This can be attributed to the lack of mauve beads near the solvated side chains on the top center and bottom left. However, for $n_k = 18$ or 10 degree resolution on the right, the AA and CG surfaces are nearly identical in Fig. 1.

The parameter n_k is used for each of the five angles representing the orientations of two rigid bodies. These five angles include the polar and azimuthal angles of the vector connecting the center of masses of the two rigid bodies and the three Euler angles of the second rigid body, with the first rigid body always held at a fixed reference orientation. If the two rigid bodies are identical, symmetry removes half of the relative orientations, and the total number of orientations is $(n_k + 1)^3(2n_k + 1)^2$. Otherwise, if the two particles differ, the total number of orientations is $(n_k + 1)^2(2n_k + 1)^3$. In both cases, a fraction of these orientations are duplicates. For example, consider when the polar angle in spherical coordinates is zero. In this case, all possible values of the azimuthal angle result in an identical separation vector. Duplicate orientations also occur in Euler angles.

A brute-force numerical method was used to determine unique orientations by verifying if the coordinates in a rigid body with the given relative orientation are identical to an existing set of coordinates. The computational cost for determining unique orientations is negligible because, for a given n_k, the process does not need to be repeated. The use of quaternions may also alleviate this issue. Figure 2 shows the total number of orientations and the fraction of those orientations that are unique as a function of n_k, the number of orientations per 180°. The process of determining unique orientations is implemented in the first step of a Free Energy and Advanced Sampling Simulation Toolkit (FEASST) version 0.25.1 tutorial in the `aniso` plugin.[79,80]

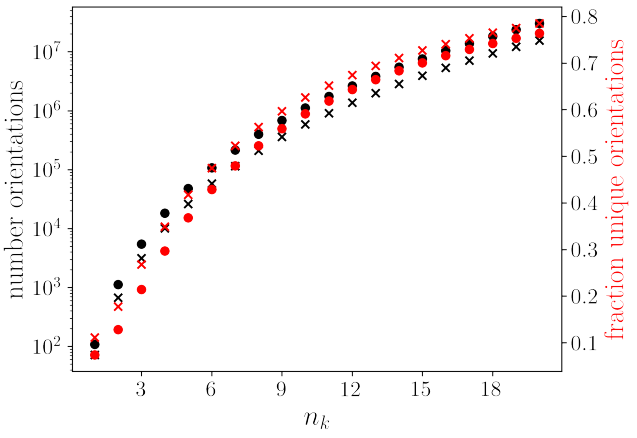

FIG. 2: The total number of orientations as a function of n_k, the number of orientations per 180°, is shown in black for (circle) interactions between identical pairs and (x) interactions between different pairs. The fraction of unique orientations is shown in red.

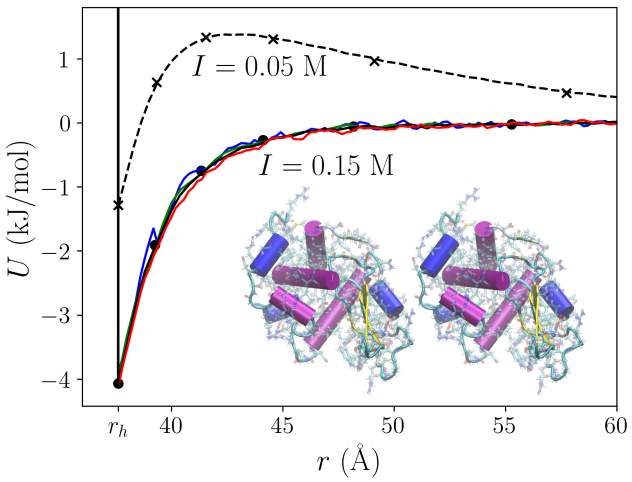

FIG. 3: The interaction energy, U, between two rigid lysozymes (PDB 4LYT at pH 6) as a function of separation distance, r, at a fixed orientation. The solid blue line is for no smoothing; green is 1 Å smoothing, black is 2 Å smoothing, and red is 4 Å smoothing with $I = 0.15$ M. The black circles are for $I = 0.15$ M (with $n_z = 7$ defined in Sec. II C). The black dashed line and the x markers are similar as above but for $I = 0.05$ M.

The choice of linearly spaced angles for approximately globular rigid bodies leads to greater resolution near the poles and lower resolution at the equator, as illustrated by the beads in Fig. 1. Future improvements to this method may incorporate spacing that is more uniform on the surface of a sphere. More spherically uniform representations may reduce the cost of the generation of the table by requiring less pre-computed orientations for a given minimum resolution near the equator, but may also increase the interpolation cost during CG simulations relative to linear interpolation. In this article, we use linear spacing to enable fast interpolation during the CG simulations. While we do not compare and contrast different interpolation methods in this article, it is also worth noting that the increased resolution at the pole could be used to orient the rigid bodies in such a way as to increase the resolution at a particularly important site such as a binding site, or a location where the rigid body is very non-spherical. Furthermore, CG potentials pre-computed non-uniformly are more an issue of efficiency than correctness, as opposed to other cases such as the violation of detailed balance in the generation of proposed MC trial orientations.

B. Atomistic (AA) implicit solvent model

This subsection concerns the interaction energy, U, between two rigid protein structures using AA models for a given orientation. The AA interaction was previously parameterized with excluded volume van der Waals and screened charge interactions in an implicit solvent fit to B_2 data.[64] Section II C, CG models are benchmarked against AA and experimental B_2.

The implicit solvent model used in this work was inspired by the work of Qin and Zhou.[64] The excluded volume and van der Waals interaction parameters are given by Autodock4,[81] while the charges are from parameters for solvation energy (PARSE).[82] Using this model, the charge screening, determined by the ionic strength, I, is approximated by the Debye length,

$$1/\kappa = \sqrt{\frac{\epsilon_r \epsilon_0 k_B T}{2 I e^2}}, \tag{1}$$

where I is the ionic strength, e is the elementary charge, ϵ_r is the dielectric constant, ϵ_0 is the vacuum electric permittivity, k_B is the Boltzmann constant, and T is the temperature.

The van der Waals interaction term was multiplied by a single fitting parameter,[64]

$$v_s = \frac{0.18(0.27 M_w + 80)}{M_w + 80}, \tag{2}$$

where M_w is the molecular weight in kilodaltons. This parameter was fit previously for five different proteins over a pH and ionic strength range.[64] In this article, we utilize this AA model with a modified cutoff.

We modified the AA interaction cutoff for the following reasons. In principle, MSMC[75] does not require a cutoff of AA interactions. A cutoff is required because interactions must be precomputed to a finite distance for storage and interpolation. In this work, the atomic cutoff distance $r_c = 5/\kappa$ is five Debye lengths.

Interaction energies for a given orientation were obtained as follows. The pdb2pqr30[83] software converted PDB coordinate files to PDB with charge and radius

(PQR) files with PARSE[82] charges and Protein pK_a (PROPKA)[84] protonation states for a given pH after removing water molecules from the PDB file. The PQR file was then converted to an input file for the FEASST MC simulation software[79] to obtain the energy as a function of relative orientation and separation distance. This process is implemented in the `write_fstprt` function in the `coarse_grain_pdb` module of `pyfeasst`, where `this typeset` is used through the article to identify software-specific names.

Fig. 3 illustrates the interaction energy between two rigid lysozymes using this AA potential for two different ionic strengths. The orientation of lysozyme is shown in the inset and given by the spherical angles $(0, 0)$ and Euler angles $(-180,\ 0,\ -180)$ degrees with the 4LYT PDB structure as the reference frame, which results in a relative displacement along the z-axis. The Euler angles are given in the proper, active intrinsic z-x-z or "x-convention."

The blue line in Fig. 3 contains no smoothing on the atomic cutoff distance, and the apparent noise in the curve is due to charges with significant contributions at r_c. The noise is significantly reduced when interactions are smoothly interpolated to zero between distances of $r_c - r_s$ and r_c, where r_s is the smoothing distance. Fig. 3 shows various smoothing distances, where $r_s = 2$ Å was utilized for the CG models in the remainder of this work. Larger or smaller values of r_s lead to more noise in Fig. 3.

The effect of the ionic strength, I, on charge screening is also shown by the black solid and dashed lines in Fig. 3 for $I = 0.15$ and $I = 0.05$ M, respectively, where the unit M is defined as mol NaCl per L. With more salt, $I = 0.15$ M NaCl in the solution, the charges are well-screened, and the interaction energy is dominated by a van der Waals well that smoothly tends to zero at a larger distance. However, with less salt, $I = 0.05$ M NaCl, charge repulsions increase interaction energy to positive values, depending upon the orientation and protein of interest. These models include short-range attraction and long-range repulsion that play significant roles in the stability of protein solutions.[48,85,86]

C. Distance resolution

The CG model distance resolution is determined by the parameter n_z. Assuming smoothly rounded surfaces with mostly positive or zero curvature, the distance at which a given orientation leads to excluded volume overlap is defined as r_h. In this case, distances smaller than r_h have infinite energy, while the interaction energy for distances greater than r_h is computed using the van der Waals and charged terms described earlier.

The contact distance, r_h, was found using a minimization algorithm as implemented in FEASST version 0.24.5[79] `GoldenSearch`.[87] The objective function of distance was defined as follows. When hard particle overlap was detected for a given distance, the objective function returned a large number divided by distance to favor larger distances for the next guess. When there was no overlap, the objective function was set to the distance to favor smaller distances in the next guess. The minimization used a tolerance of 10^{-4} Å. Once the minimum of the objective function was found, twice the tolerance was added to avoid numerical precision issues (e.g., a new energy calculation at precisely r_h could lead to overlap). The contact distance calculation is implemented in the second step of a FEASST version 0.25.1 tutorial in the `aniso` plugin.

Interactions beyond distances of $r_h + r_c$ are zero, where r_c is the atom-based interaction cutoff distance. We linearly interpolate the interactions to zero at the cutoff over a short distance of r_s. For this reason, interaction energies are tabulated between r_h and $r_h + r_c - r_s$. Here, we transform this distance as a parameter that varies from $z = 0$ at r_h to $z = 1$ at $r_h + r_c - r_s$,

$$z = \frac{r^\gamma - r_h^\gamma}{(r_h + r_c - r_s)^\gamma - r_h^\gamma}, \tag{3}$$

where $\gamma = -4$ is a stretching exponential parameter that increases resolution for smaller distances when negative.[59,60] The distance resolution is given by n_z, the number of distances that linearly span from $z = 0$ to 1. In Fig. 3, the symbols show the resolution $n_z = 7$ with $\gamma = -4$. During the CG simulations, values for z between the n_z values were linearly interpolated. The energy precalculation calculation is implemented in the third step of a FEASST tutorial in the `aniso` plugin.

Assuming a single r_h for a given relative orientation might lead to issues in appropriately accounting for configurations where the two proteins form a hook, latch, or handshake configuration (e.g., multiple infinite maximums when infinite energies occur at distances that are between finite energy distances). These configurations are infrequent for the rigid bodies considered in this work. In the rare case U was highly repulsive for $z > 0$, and U was set to 5 kJ/mol, so linear interpolation did not lead to overlap-like energies in the entire region above and below this value of z, as implemented in the FEASST version 0.24.5 with the `max_energy_set` keyword in the class `TabulateTwoRigidBody3D`.[79]

III. SECOND OSMOTIC VIRIAL COEFFICIENTS (B_2) OF SINGLE RIGID BODIES

In Section II, the CG model of lysozyme was defined as a single rigid body represented by one anisotropic interaction site. This section uses MSMC to compute B_2 as benchmarks to compare the CG models of various resolutions against the AA values and experimental results. The MSMC B_2 calculation is implemented in the fourth step of a FEASST version 0.25.1 tutorial in the `aniso` plugin. Although B_2 may be computed without MSMC either by directly integrating the precomputed interactions or using an FFT approach,[64] we did not systemati-

cally compare the efficiencies of these different methods. An example of integration of the precomputed interactions can be found in the `VisitModelInnerTable` class of FEASST or the `anglescan` example in the open source software named Faunus.[88]

A. Mayer-sampling Monte Carlo

MSMC[75] computed B_2 using FEASST version 0.24.5.[79] Because the models use implicit solvent, the methodology for osmotic B_2 is identical to that of gas phase B_2 calculations, where the solvent degrees of freedom are integrated into the potential model and, therefore, are not sampled during MSMC. As opposed to Metropolis MC, MSMC only requires two bodies to compute B_2, which is performed at infinite dilution and without the need for periodic boundary conditions in the implicit solvent. In addition, excluded volume overlap is allowed in MSMC and is an essential part of the sampling technique, which aims to use MC integration to obtain the Mayer function divided by the Mayer function of a reference potential. The reference potential is often chosen as a hard sphere with known B_2 and of appropriate size that is well sampled during the MSMC simulation with the full potential, although other choices are also possible.[89]

To begin MSMC, two rigid bodies were placed in overlapping identical positions and orientations. Trial moves include translation of only the second rigid body or rotations of either rigid body. The MSMC reference potential was a hard sphere of diameter σ_r at the center of mass of the rigid body. The maximum translation was optimized 10^3 times during an equilibration period every 10^5 trials to reach a target acceptance of 25% Trial rotations were generated by perturbing quaternions with a weighted random quaternion,[90] and this weight was optimized in the same fashion as the maximum translation. The simulations were terminated after 3×10^9 trials for CG lysozyme, which required 5 days of central processing unit (CPU) clock time on high performance compute nodes with dual Intel® Xeon® Silver 4216 CPUs with a total of 32 processors per node at 2.10 GHz base frequency (see disclaimer at the end of Section VII). To obtain the standard error of the mean, every condition and parameter set was simulated independently on each of the 32 processors with different random number seeds. For the slower AA simulations, maximum displacement and rotation parameters were tuned during an initial equilibration period 10^2 times every 10^2 trials, with over 1.5×10^6 trials for lysozyme. For the even larger lactoferrin (PDB 1BLF), MSMC with 1.5×10^5 trials were run for 2 weeks. The σ_r was 30 and 50 Å for lysozyme and lactoferrin, respectively. The B_2 values were reported in units of 10^{-4} mol ml g^{-2} using molecular weights from the PQR files described in Section II B.

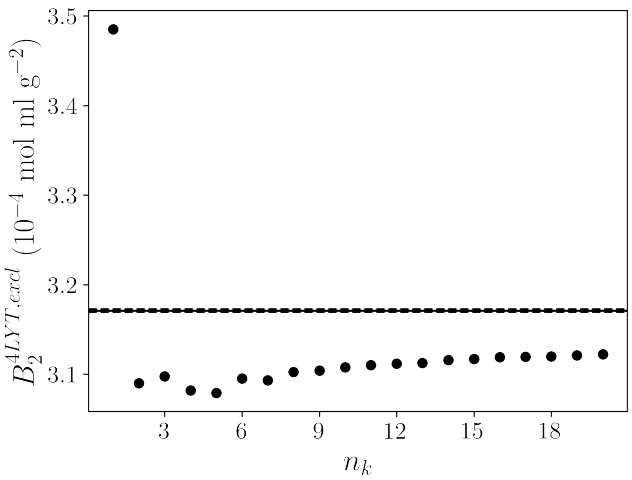

FIG. 4: The B_2 of lysozyme (PDB ID 4LYT) with only excluded volume interactions as a function of the number of orientations per 180°, n_k for the CG model. Error bars are the standard error of the mean and are smaller than the symbols. The solid horizontal line shows the average of the AA model value, while the dashed lines outline the range of uncertaintly from the standard error of the mean.

B. B_2 of AA and CG lysozyme

In Subsection III A, MSMC simulations were used to compute B_2 for both lysozyme and lactoferrin using AA models and CG models that assumed the entire proteins were a single rigid body in their PDB structure with an implicit solvent AA potential model as described in Section II B. Figure 4 shows the excluded-volume B_2 of lysozyme (4LYT), $B_2^{4LYT,excl}$ for the AA model and the CG model as a function of the angular resolution, n_k, which is the number of angles per 180° in each of the five relative orientation angles. Only hard particle interactions are included in the excluded-volume $B_2^{4LYT,excl}$; van der Waals and screened charge terms are omitted. For the AA model, $B_2 = 3.171 \pm 0.002$ (10^{-4} mol ml g^{-2}), where $\pm$ is the standard error of the mean from 32 independent simulations. As shown in Fig 4, as n_k increases, $B_2^{4LYT,excl}$ reaches a plateau. The B_2 of CG simulations approach the AA B_2 as angular resolution, n_k increases but is still offset at the largest values of n_k. This may be attributed to the assumption that r_h is well defined for a given orientation, which is invalidated by hook, latch and handshake configurations. Overall, the excluded-volume B_2 of the CG model agrees well with that of the AA model.

Because the excluded volume interactions are represented by a single distance, r_h, for a given relative orientation, the distance resolution, n_z, plays no role in the calculation of $B_2^{4LYT,excl}$. Therefore, $B_2^{4LYT,excl}$ was used to determine a reasonable value of n_k that represents the shape. We select $n_k = 12$ as a reasonable trade-off

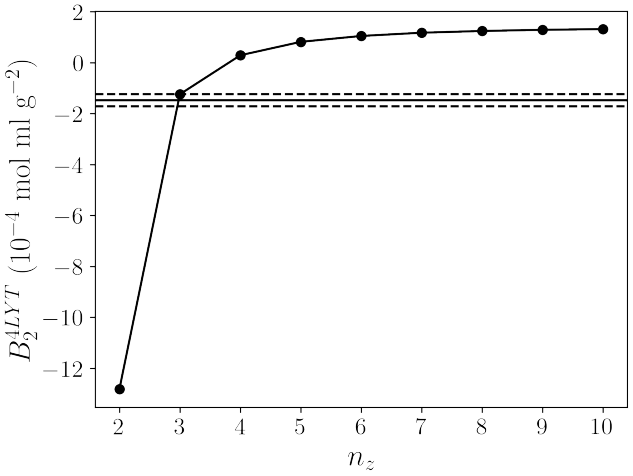

FIG. 5: The B_2 of lysozyme (PDB ID 4LYT at pH 6) as a function of the number of distances, n_z, per orientation with $n_k = 12$. Error bars are the standard error of the mean and are smaller than the symbols. The solid horizontal line shows the average of the AA model value, while the dashed lines show ± the standard error of the mean.

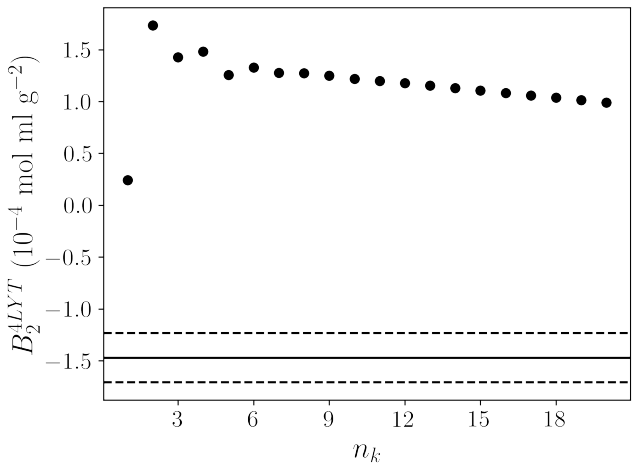

FIG. 6: The B_2 of lysozyme (PD ID 4LYT at pH 6) as a function of the number of orientations per 180°, n_k, for $n_z = 7$. Error bars are the standard error of the mean and are smaller than the symbols. The solid horizontal line shows the average of the AA model value, while the dashed lines show ± the standard error of the mean.

between the accuracy shown in Figs. 1 and 4 and the computational cost of increasing n_k implied in Fig. 2.

The B_2 of lysozyme (PDB ID 4LYT), B_2^{4LYT}, with all interactions including van der Waals and screened charges is shown in Fig. 5 for both the AA model and the CG models with varying distance resolution, n_z, at fixed orientational resolution, n_k. For low n_z, attractions are overestimated due to the shape of the short-range attraction, as shown in Fig. 3. This leads to a fortuitous cancellation of errors when $n_z = 3$, resulting in a prediction close to the AA value. However, as n_z increases, B_2^{4LYT} reaches a plateau that is shifted from the AA value. One possible explanation for why B_2^{4LYT} is shifted from the AA value is because of the fixed orientational resolution of $n_k = 12$. In order for the CG model to converge to the AA value with increasing resolution, both n_k and n_z would need to be increased. For the remainder of this work, we use $n_z = 7$ as a reasonable trade-off between accuracy and computational cost. We omit $n_z = 3$ because the fortuitous cancellation of errors cannot be guaranteed for other proteins or other solution conditions. If we wish to match the virial coefficient of the CG models to the AA model, we can do so by scaling the interactions without reducing the resolution of the CG model. Instead, $n_z = 7$ shows a reasonable level of convergence to the plateau value.

With $n_z = 7$ chosen as a reasonable distance resolution, we now revisit the orientational resolution of the full potential. Figure 6 shows the B_2^{4LYT} as a function of n_k, the number of orientations per 180°, for $n_z = 7$. Similar to Fig. 4, the CG model slowly moves toward the AA value as n_k increases after an initial jump at very low resolution. By coincidence or fortuitous cancellation of errors, $n_k = 1$ is the closest to the AA value. Otherwise, the attractions steadily increase with n_k toward the AA value. Similar to Fig. 5, the CG model does not quite reach the AA value, likely because the distance resolution, n_z, is fixed, in addition to the other assumptions discussed for the results in Figs. 4 and 5.

Computational costs make investigating higher values of n_k and n_z impractical. Generating such a large precomputed table from AA interactions becomes very costly. When the tables become large, interpolation during the simulation can also slow down because the CPU cache is less likely to have useful information for the next query. A simulation typically should not require more than 1 or 2 gigabytes of random access memory (RAM) per processor on modern high-performance computer clusters. To further reduce memory requirements, the table was stored in single precision.

For the values of n_k studied up to the limit of practicality, the CG model has fewer attractions than the AA model. We hypothesize that the limited number of orientations may miss some important attractive patchy interactions, which contribute to the virial coefficient. Previous work in the molecular surface recognition and docking literature can reach higher resolutions[91] that are feasible when only the most attractive orientations are stored and highly optimized methods are used in the orientational search.[63]

Fig. 7 compares the CG and AA models with experimental data for lysozyme at pH of 4.5 for various ionic strengths. Experimental data were estimated using WebPlotDigitizer[97] for illustrative purposes only. There is some variation in the experimental data, which could be due to measurement uncertainty or possibly different

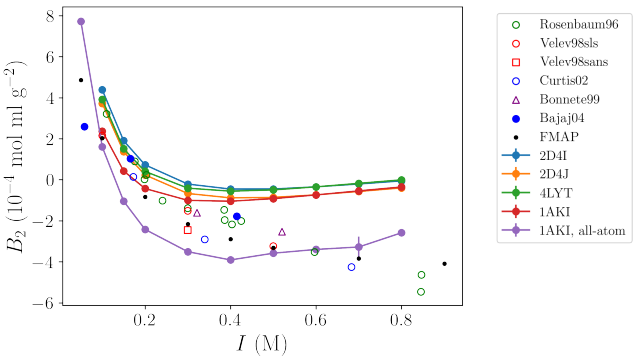

FIG. 7: The B_2 of lysozyme at pH 4.5 from previously published experimental data[92–96] is shown with open symbols and the blue circle. The CG models for various PDB IDs (4LYT, 1AKI, 2D4I, 2D4J) as a function of the ionic strength, I, with $n_k = 12$, $n_z = 7$, are shown by the lines. AA values for 1AKI are shown in purple and the FMAP[64] results with a different protonation state are also shown by the black dots. Error bars are the standard error of the mean and may be smaller than the symbols.

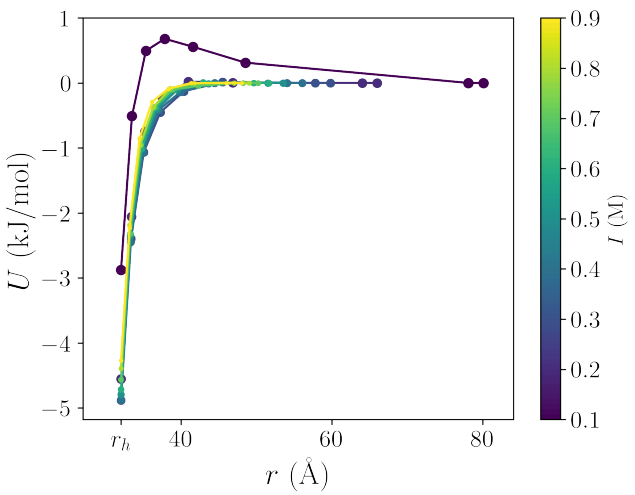

FIG. 8: The potential energy of the 1AKI CG model as a function of separation distance for nine values of I at a fixed orientation where all angles are their maximum values.

formulation conditions, such as different buffers or other technical details of the instruments. The previous computational results using the FFT-based method for Modeling Atomistic Proteins–crowder interactions (FMAP) for computing B_2 by Qin *et al.* are also shown in Fig. 7.[64] Their AA results differ from ours because for the special case of lysozyme, Qin *et al.* used experimental pK_a determined at a temperature of 25° C and 0.1 M ion strength, while this work used PROPKA[84] assignments from pdb2pqr30.[83] Although the approach used by Qin *et al.* agrees more with the experiment over a wide range of I; we treat all of the proteins with the same methodology that is easy to apply to the NISTmAb or other proteins of interest.

One major difference between the previously published experimental and simulation results and the results in this work is the increase in B_2 as ionic strength increases beyond 0.4 M. This increase of B_2 is a consequence of the AA model described in Section II B. Fig. 8 shows the potential energy as a function of protein center-of-mass separation distance at a fixed orientation for various ionic strengths. At the lowest $I = 0.1$ M, there is a short-range attraction and a long-range repulsion. As I increases, the charges are more screened, and the long-range repulsion no longer results in positive potential energy values at intermediate separation distances. In addition, U at contact decreases with increasing I until it reaches a minimum near $I = 0.3$ M and then gradually increases to less attractive values. From $I = 0.3$ to 0.9 M, the attractive well width also decreases. This could explain the increase in B_2 with I above 0.3 M shown in Fig. 7, which is not seen in the experiments. Another explanation is that the approximation of Eq. 1 breaks down for higher values of I.[98] Because of the different titration treatments for pH, the difference between our results and Qin *et al.* is likely due to various charge interactions.

C. B_2 of AA and CG lactoferrin

We will now end our discussion of B_2 for models that assume the entire protein is a rigid object by considering a different protein. Lactoferrin is known to form a dimer at specific ionic strengths.[99] Therefore, lactoferrin is an interesting test case to see if the CG methodology captures specific, short-range, patchy, attractive interactions. Lactoferrin is larger than lysozyme with a molecular weight of over 75 kDa. Therefore, lactoferrin tests the limits of the computational costs of the CG model. Lactoferrin is also less spherical than lysozyme and may be better represented as two bonded rigid objects rather than one. Regardless, we consider lactoferrin to be a single rigid object in this work.

The B_2 of lactoferrin is shown in Fig. 9. The experiments show a minimum in B_2 as a function of ionic strength, which was previously associated with the formation of dimers.[99–102] The AA and CG models also show a minimum, although the minimum occurs at a larger ionic strength of approximately $I = 0.1$ M compared with experiments. The difference between the experiment and the AA model is likely due to a number of assumptions. These assumptions include the usage of the rigid crystal structure, the implicit-solvent model, approximate van der Waals and charge-screened terms, and protonation state assignments. The specific dimer interaction was also difficult for the MSMC simulations to sample. This led to very large error bars around the minimum in B_2 at $I = 0.1$ M because some of the 32 in-

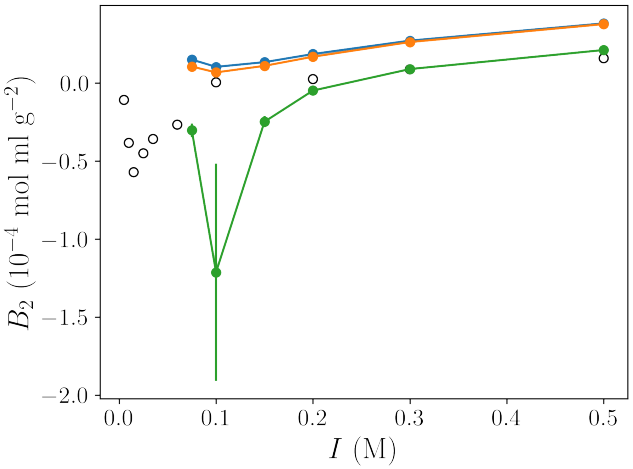

FIG. 9: The B_2 of lactoferrin (PDB ID 1BLF) at pH 7 with previously published experimental data shown by the open circles,[100] and the CG models as a function of the ionic strength for different orientational resolution, (blue) $n_k = 10$ and (orange) $n_k = 12$. AA values are shown in green. Error bars are the standard error of the mean and may be smaller than the symbols.

dependent simulations could find the specific dimer configuration while others were not. Even though a single MSMC simulation was run for two weeks with 32 independent replicas, the number of trials was relatively limited due to the computational expense of this large protein.

The CG models are also more repulsive than the AA model. This is likely because the CG model misses important specific interactions via limited orientational resolution. Due to the size of lactoferrin, the same n_k as lysozyme represents a lower orientational resolution because the arc length would be larger. This is also likely why a lower resolution CG model with $n_k = 10$ has a higher and more repulsive B_2 than $n_k = 12$.

IV. MULTISCALE ANTIBODY MODELS

In Section II, we modeled the entire lysozyme and lactoferrin protein as a single anisotropic site with interactions precomputed from a rigid body given the PDB atomic structure. In this section, we develop a model for the NISTmAb, where the antibody was modeled as a collection of three rigid bodies bonded to one another with a flexible linker or hinge. The bonded potential was parameterized from microsecond-long AA MD simulations with explicit solvent. We then apply MSMC to the flexible CG model to compute B_2. For the NISTmAb, we compare the B_2 against experimental data.

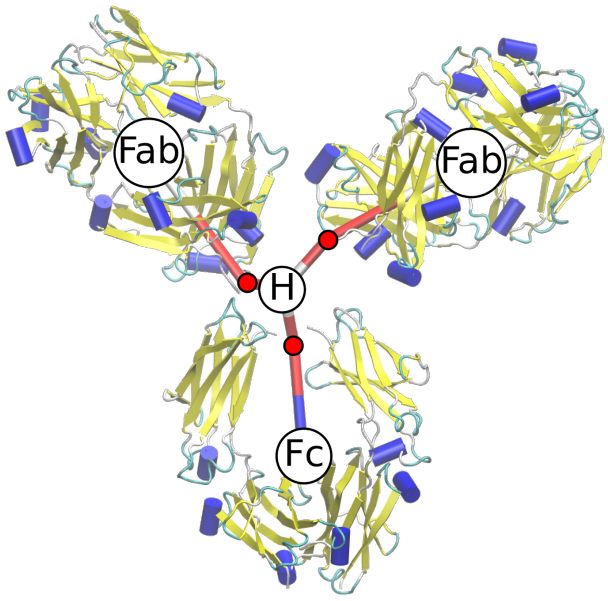

FIG. 10: Illustration of the CG model of the NISTmAb. The center of mass of the three rigid domains and hinge are labeled, while the backbones of the three rigid domains are shown by their ribbon representation using VMD.[76] Each rigid domain connects to the hinge at the red circles, which form elbows. The angle of the elbow and the length between the hinge and elbow are both highly flexible and parameterized with NISTmAb AA molecular dynamic simulations with explicit solvent.

A. The coarse-grained (CG) antibody as three bound rigid bodies

As shown in Fig. 10, an antibody is made of three relatively rigid domains connected by a relatively flexible hinge region.[52] The crystallizable fragment (Fc) comprises approximately half of two highly conserved heavy chains, and the antigen-binding fragment (Fab) comprises the other approximately half of a heavy chain combined with a light chain. The rigid Fc and Fab domains are labeled by their center of mass in Fig. 10. For the NISTmAb, the two heavy and light chains and the two Fabs are identical.

The flexible hinge region is made of the middle section of the two heavy chains. Cysteine residues form disulfide bonds between the heavy chains, and proline residues limit backbone rotation relative to other residues. The center of the hinge was defined as the center-of-mass of the cysteine-proline-proline-cysteine (CPPC) residues and labeled as the H in Fig. 10. Specifically, for the NISTmAb PDB,[103] this corresponds to the CYS, PRO, PRO, CYS residues 229-231 of both heavy chains. The point where the hinge is connected to each of the three domains is shown as a red circle in Fig. 10. These three connections form elbows, where the bond between the

elbow and the center of each domain is held rigid. In contrast, the bond between these three connections and the hinge is highly flexible. The bond length between the hinge and the three connections is allowed to change, and the bond angle with the connections as the vertex between the hinge and each of the three rigid domains. These bonds will be parameterized by comparison with microsecond-long AA MD simulations of a single NISTmAb in an explicit solvent.

The connections between the hinge and the three rigid domains are defined as follows. For the rigid Fab domains, the connection is defined as the center of mass of the two residues in the heavy and light chains that form the disulfide bond near the hinge. Specifically for the NISTmAb PDB,[103] this corresponds to the CYS residue 223 on heavy chain H and residue 213 on light chain L. For the rigid Fc domain, the connection to the hinge is defined as the center of mass of the proline residue on both heavy chains just after the two glycine residues that are part of the flexible hinge. Specifically for the NISTmAb PDB,[103] this corresponds to the PRO residue 241 on heavy chains H and V.

B. All-atom (AA) explicit solvent MD simulations

AA MD simulations of the NISTmAb with explicit solvent were used to inform the flexibility of the CG models. In particular, bonded potentials for the flexible bond angles and lengths where the three rigid domains connect to the hinge are unknown, as shown in Fig. 10. In this subsection, we use AA MD simulations of the NISTmAb to investigate the minimum and maximum bond lengths and minimum bond angle observed as the antibody is allowed to move freely in solution.

The AA MD simulations were started from two conformations: a "hands-up" Y-shaped model and the atomic model of the NISTmAb, with an alternative hinge conformation that better matched deposited crystal structures of intact mAbs.[103–105] The automated tool `prepareforleap`[106] was used to map glycans, determine cysteine crosslinks, and set amino acid protonation states assuming a pH of 7. Asiaylated, bi-antennary core fucosylated complex (G0F and G1F) glycoforms were included and were based on crystal structure coordinates from the NISTmAb Fc domain.[104] These glycoforms are consistent with the most abundant glycoform class identified via mass spectrometry.[107] Structures were built using tleap in AmberMD.[108] The ff14SB protein force field,[109] glycam 06j-1 glycan force field,[110] extended simple point charge (SPC/E) water model,[111] and Joung-Cheatham monovalent ion force field[112] were used to parameterize the system. To reproduce multiple experimental ion concentrations, the structure of NISTmAb was simulated in either neutralizing conditions (ten Cl^- counterions), 0.150, 0.5 or 1 M NaCl (Table I). Ion positions were randomized by swapping coordinates with water molecules, 6 Å away from the solute and 4 Å away from each other, fourtimes, and positions saved, generating four starting structures with different solvent positions for each system.

Starting structures (four per system) were minimized and equilibrated by a ten-step protocol, including successive rounds of minimization and MD with decreasing positional restraints in an isobaric-isothermal (NPT) and canonical (NVT) ensemble, as described previously.[113] Production dynamics were performed using the AmberMD engine[114] with graphics processing unit acceleration and single precision and 64-bit fixed point integer arithmetic. Solute hydrogen masses were repartitioned to 3.024 Da by reducing the mass of the heavy atom to which the hydrogen is bound, with rigid bond constraints,[115] enabling a 4 fs timestep for production dynamics.[116] A direct space cutoff of 9 Å was used for particle mesh Ewald.[117] For NPT simulations, the pressure was set to 1 atm and regulated using a MC barostat.[118] The temperature was set to 300 K and regulated by a Langevin thermostat[119] using a collision frequency of 5 ps^{-1} and setting a flag to prevent synchronization artifacts.[120] For NVT simulations, the temperature was set to 300 K and regulated by a Langevin thermostat[119] using a collision frequency of 2 ps^{-1}. Simulations were run, and trajectories were saved for either 1 μs or 500 ns. Table I contains simulation details of each system.

To parameterize the CG 7-site model, analyses were performed on 212 500 total frames covering 22 total microseconds of MD simulation. All frames from all simulations detailed earlier were used. Distances were calculated between the center-of-mass (COM) of the CPPC motif in the hinge region and the COM of two residues making up the last Fab residues (CYS-CYS) or start of the Fc domain residues (GLY-GLY) before the hinge region started. Vector math was performed in CPPTRAJ to calculate the angles between domains.[121]

The probability of the distance between where the three domains connect to the hinge and the center of the hinge is shown in Fig. 11. These probabilities were computed from the aggregate of all trajectories summarized in Table I, which include different salt concentrations, starting configurations, periodic boundaries, and ensembles. All of the data was included to find the limits of flexibility, which we will use as an input to the CG models. Because the CG model will include intramolecular domain-domain interactions that are precomputed from rigid AA structures, capturing the specific protein-protein interactions that contribute to all of the minima and maximum in the distributions shown in Fig. 11 is not the purpose of this parameterization. Instead, we want to know what reasonable minimum and maximum lengths and minimum angles are needed for hinge flexibility.

From the AA explicit solvent MD simulations shown in Fig. 11, we choose round numbers for the minimum and maximum lengths of the CG model of 8 and 25 Å, respectively, for the Fab domains and 12 and 26 Å, respectively, for the Fc domain. These choices will be further tested

TABLE I: Full-length NISTmAb AA simulation details of an aggregate 22 microseconds of MD.

I (M)	Start id	Water	Cl^-	Na^+	Atoms	Time per replica (μs)	Replicas	Ensemble
0	1	122 784	10	0	358 943	1	4	NPT
0	2	166 652	10	0	482 747	1	4	NPT
0.15	1	112 050	377	367	357 475	1	4	NVT
0.15	2	101 481	377	367	325 768	1	4	NVT
0.5	1	110 336	1234	1224	354 047	0.5	4	NVT
1	1	107 888	2458	2448	349 151	1	4	NPT

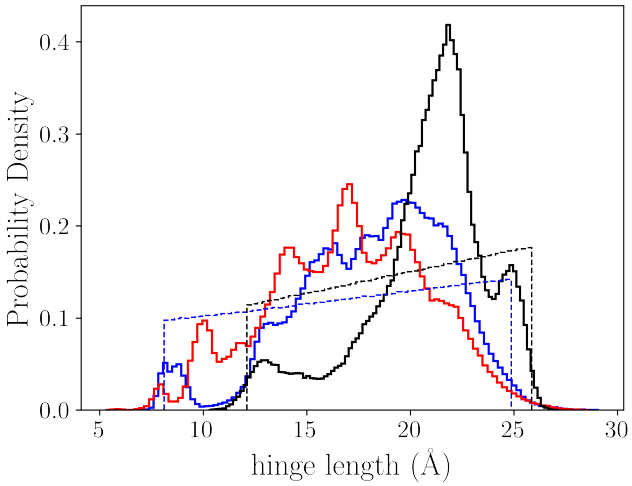

FIG. 11: The probability of the distance between the center of the hinge and where the three domains connect to the hinge, as described in Fig. 10, for (solid lines) AA explicit water MD with (blue) Fab1-H, (red) Fab2-H, (black) Fc-H, and (dashed lines) the CG model. The results for the CG model used only excluded volume interactions without van der Waals or screened charge interactions, while the results for the AA model include the full potential and span a range of salt concentrations.

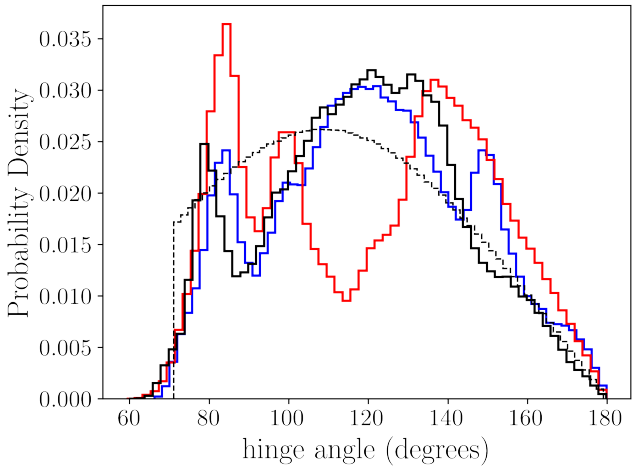

FIG. 12: The probability of the angle whose vertex is given by the red circles described in Fig. 10 is shown with the same colors as described in Fig. 11.

by comparing the distribution with CG MC simulations of a single antibody with the MD results and with experimental data for the radius of gyration of the flexible CG model.

The probability of the angle whose vertex is given by the location where the domains connect to the hinge and endpoints as the center of the hinge and each domain is shown in Fig. 12. Similar to Fig. 11, the probabilities include simulations over various solvent conditions to find the minimum angle that will be input to the CG MC simulations. Again, we choose round numbers for the minimum according to the MD simulations. In this work, we choose 70° as the minimum angle for all three domains. The maximum angle is 180°, as expected for a highly flexible hinge. The Fab1 had an angle distribution closer to Fc than Fab2, but the minimum angle used in the CG model was the same for all three domains.

C. Flexible coarse-grained model using configurational-bias MC

In Subsection IV B, AA MD simulations were performed to obtain the minimum and maximum bond lengths and minimum angle for the flexible hinge in the CG model. In this subsection, we use configurational-bias (CB) MC[122] to simulate the flexible CG model with three anisotropic interaction sites bonded by the flexible hinge region. To simulate a highly flexible antibody hinge, the bonded potential energy is zero when the bond lengths and angles are within the allowed range and infinite when outside of this range.

Each CB trial proceeds as implemented in FEASST version 0.24.5 `BondSquareWell` and `AngleSquareWell` potentials with `TrialGrow` bond and `rigid_body_angle` keywords.[79] Randomly select one of the three rigid domains. Place the connector of that domain randomly in a spherical shell within the allowed bond distances. Randomly choose an angle according to the fully flexible sine distribution that is within the allowed angle range. Place the center of the domain uniformly randomly about the circle of allowed positions given the angle. Accept or reject the trial according to the change in energy that includes intramolecular and intermolecular interactions.

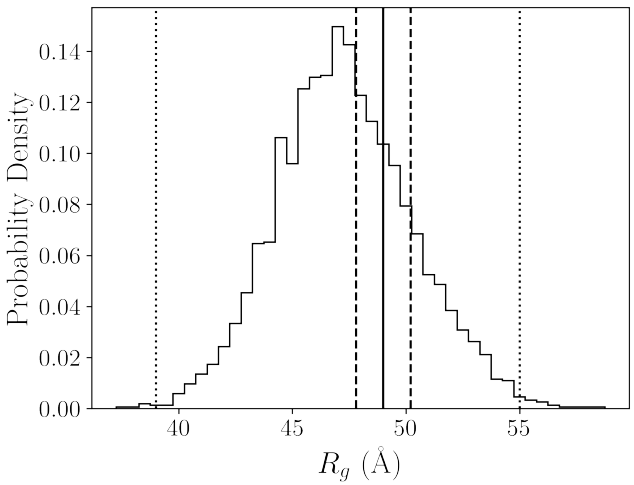

FIG. 13: The probability of the radius of gyration of the CG model of the NISTmAb for pH= 6, $I = 0.15$ M, $n_k = 12$, and $z = 7$ is shown with the solid lines. The solid vertical line is the experimental value of 49 Å, and the dashed lines show the experimental standard deviation of 1.2 Å.[123,124] The dotted vertical lines correspond to the lower and upper limits of 39 and 55 Å from previous MC methods.[125] The average from the CG simulations in this work was 47.31 Å with a standard error of the mean of 0.05 Å.

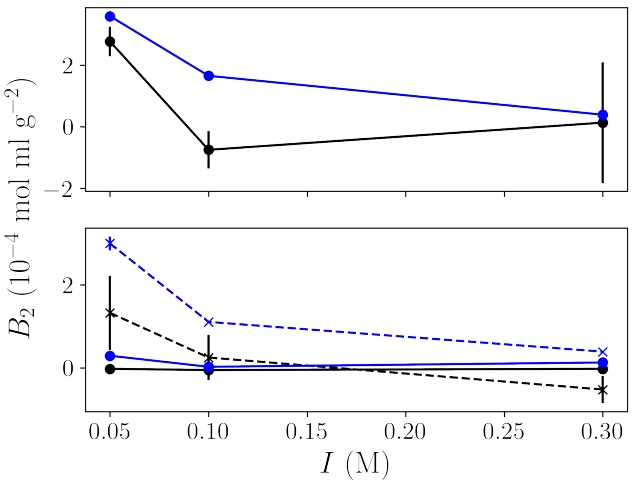

FIG. 14: (Top) Second virial coefficient, B_2 of the Fab domain of the NISTmAb from (black) previous scattering experiments[67] and (blue) MSMC with a CG model, $n_k = 12$, $n_z = 7$, pH 6. (Bottom) B_2 of the entire NISTmAb for (solid lines with a circle) pH 6 and (dashed lines with x) pH 3.

Figs. 11 and 12 show the resulting bond and angle distributions, respectively, for a CG MC simulation of a single NISTmAb with only excluded volume interactions and no van der Waals or screened charge interactions. Only excluded volume interactions were included to illustrate the expected distributions for a highly flexible antibody. These bond potentials should be independent from the specific Fc and Fab domain interactions for reuse with other mAbs with the same hinge structure. Despite omitting highly specific interactions between the domains, the overall shape of the bond and angle distributions are represented with such a model. For example, the bond distance probability is expected to grow with distance according to random point picking in a spherical shell. In addition, the bond angle probability is expected to reach a maximum according to the expected sine distribution for random point picking on the surface of a sphere. Although specific features appear in the distribution due to specific interactions between the domains in the MD simulations, the CG model aims to capture these effects using the precomputed AA potentials.

The flexible potential for the CG model is validated by comparison of the radius of gyration with experimental data, as shown in Fig. 13. A MC simulation of a single CG NISTmAb was performed in a 496 Å periodic box with 3×10^9 CB trials that include the full intramolecular potential with excluded volume, van der Waals and screened charges with solution conditions of pH= 6 and $I = 0.15$ M, and CG resolutions of $n_k = 12$ and $n_z = 7$. The radius of gyration was computed by back-mapping the atomic positions of the rigid domains given the anisotropic site orientations, while all atoms were assumed to have the same mass.

D. Second osmotic virial from coarse-grained MSMC with CB

In this subsection, MSMC with CB of two proteins is used to compute B_2 of both the Fab domain and the flexible CG model of the NISTmAb shown in Fig. 10. The B_2 values of the NISTmAb Fab domain, shown in the top panel of Fig. 14 for pH 6, were computed using the methodology described in Section III A for a single rigid domain, with over 4×10^9 trials. The σ_r values were 45 and 85 Å for the Fab and NISTmAb, respectively. The B_2 values of the NISTmAb Fab follow similar trends in ionic strength dependence as the experimental values but are more repulsive than observed in the experiment. This is likely due to the CG model missing specific attractive interactions at short range due to the limited resolution in orientation and distance. Note that the experimental values depend on the buffer, and the effect of different buffers is not captured in our implicit solvent AA or CG models. The B_2 of the CG model of the NISTmAb Fab compared best with the experiment around $I = 0.3$ M.

For MSMC with a flexible CG model, rigid translations and rotations were utilized as described in Section III A, as well as a flexible CB trial that is different from the canonical ensemble as described in Section IV C. There are two stages for acceptance and rejection of a trial. The first stage is to find a new configuration that is

acceptable according to the intramolecular interactions. The second stage is to accept or reject the trial using standard MSMC acceptance based on the intermolecular interactions.[52,126] In MSMC, each CB attempt proceeds as follows. One of the three domains is partially regrown as described for the single-mAb canonical MC, and this regrowth is repeated as many times as necessary until the proposed configuration is accepted according to the Boltzmann-weighted change in the intramolecular interaction energy. Once a newly proposed internal configuration of one of the mAbs is accepted, the trial move is accepted or rejected using the standard MSMC acceptance where the energy of interaction is given only by the intermolecular interactions.[52,126]

The bottom panel of Fig. 14 shows the B_2 values of the flexible NISTmAb CG model compared to experiments over a range of ionic strengths and pH 3 and 6. For pH 3, B_2 is shifted by nearly a constant factor, where the CG model has more repulsions, as also seen by comparison with the Fab domains. For pH 6, B_2 is again shifted by a positive constant factor relative to the experiment, but in this case the CG B_2 is much closer to the experiment than for pH 3. For more accurate predictions with CG CB MC simulations of hundreds of NISTmAb, the CG potential can be scaled by fitting to match the experimental B_2 values. Such a fit is expected to correct for the overall average interaction strength while retaining the relative anisotropic interactions.

V. CONCLUSIONS

We demonstrate a CG methodology where rigid all-atom (AA) models with implicit solvent are precomputed, stored, and interpolated during MC simulations. The computational results were compared against experimental second osmotic virial coefficients (B_2) and the radius of gyration over a variety of angular and distance resolutions. Many of the CG modeling assumptions could be relaxed in future work at the cost of more computational effort, including explicit solvent, flexibility of side chains, better AA models, and improved resolution and interpolation.

In principle, representing a protein as an increasing number of bound rigid bodies may improve the CG model by relaxing the major approximation at the expense of computational cost. For example, proteins can be divided into parts based on their secondary structure or amino acids.[43] Such a process requires parameterization for bonded potentials between rigid bodies. Normal mode analysis could identify relatively rigid subunits and the motions between these domains.[127] If a domain was instead described by two or more rigid structures, these structures could be pre-computed, and a multi-conformation MC trial[55] could be used.

One explanation for the difference between the B_2 of the AA and CG models is that linear interpolation may capture attractive patches inaccurately. On the other hand, interpolation errors may be comparable to or smaller than the assumptions involved in using implicit solvent models. Comparisons of CG results to experimental data are also a meaningful test of accuracy. The AA fitting parameter for the short range attractive van der Waals attractions could be specifically parameterized for the CG model to fit AA or experimental data. Future work on AA parameterization could lead to improved comparison between CG models and experiments. Future improvements to storing and interpolating may allow the tabular values to better converge toward AA values.

Another assumption that could be improved is that the solvent may be represented explicitly. The use of explicit solvents would greatly increase the computational expense of this method compared to the implicit solvent models used in this article because the ensemble averages of the interaction energies between the rigid bodies would need to be computed for each orientation to obtain the potentials of mean force. With enough computer resources, other solvation models may be possible.[128] Enabling explicit solvents could also be one way of incorporating co-solvents and co-solutes. On the other hand, solutes and formulation excipients may also be modeled by precomputing and tabulating their interactions with the rigid domains and themselves for explicit representation in the CG model. While this study focuses on biomolecules, the methodology could be applied to arbitrarily complex models.

Although this article focuses on demonstrating the CG methodology and validation with low-concentration measurements such as B_2 and the radius of gyration at infinite dilution, future work could include comparisons of higher-concentration measurements such as experimental small-angle scattering. Comparisons of viscosity as a function of concentration with the experiment would also be very informative but may require non-trivial extensions of the CG method to MD.

VI. SUPPLEMENTARY ONLINE MATERIAL

The Supplementary Online Material contains data for all figures in comma-separated value format in the file "fig_data.zip."

VII. ACKNOWLEDGEMENTS AND DISCLAIMER

This article was funded by the National Institute of Standards and Technology and is not subject to U.S. Copyright. Certain commercial firms and trade names are identified to specify the usage procedures adequately for reproducibility. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that related products are necessarily the best available for the purpose.

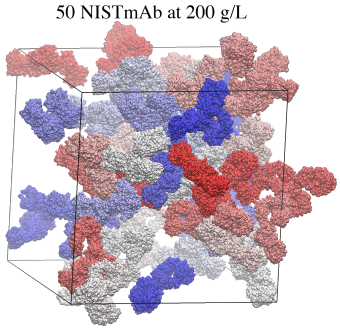

FIG. 15: Table of contents figure.

[1] A. Brown, Nat. Rev. Drug Discov **22**, 8 (2022).
[2] N. A. McGrath, M. Brichacek, and J. T. Njardarson, J. Chem. Educ. **87**, 1348 (2010).
[3] T. A. Waldmann, Nat. Med. **9**, 269 (2003).
[4] S. J. Shire, Z. Shahrokh, and J. Liu, J. Pharm. Sci. **93**, 1390 (2004).
[5] A. Matucci, A. Vultaggio, and R. Danesi, Respir. Res. **19**, 154 (2018).
[6] S. J. Shire, Curr. Opin. Biotechnol. **20**, 708 (2009).
[7] W. F. Weiss, T. M. Young, and C. J. Roberts, J. Pharm. Sci. **98**, 1246 (2009).
[8] W. Wang and C. J. Roberts, *Aggregation of Therapeutic Proteins* (John Wiley & Sons, 2010).
[9] C. J. Roberts, Trends Biotechnol. **32**, 372 (2014).
[10] A. Chaudhri, I. E. Zarraga, T. J. Kamerzell, J. P. Brandt, T. W. Patapoff, S. J. Shire, and G. A. Voth, J. Phys. Chem. B **116**, 8045 (2012).
[11] M. M. Castellanos, J. A. Snyder, M. Lee, S. Chakravarthy, N. J. Clark, A. McAuley, and J. E. Curtis, Antibodies **6**, 25 (2017).
[12] J. E. Schiel, M. J. Tarlov, K. W. Phinney, O. V. Borisov, and D. L. Davis, in *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 3. Defining the Next Generation of Analytical and Biophysical Techniques*, ACS Symposium Series, Vol. 1202 (American Chemical Society, 2015) pp. 415–431.
[13] J. P. Marino, R. G. Brinson, J. W. Hudgens, J. E. Ladner, D. T. Gallagher, E. S. Gallagher, L. W. Arbogast, and R. Y.-C. Huang, in *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 3. Defining the Next Generation of Analytical and Biophysical Techniques*, ACS Symposium Series, Vol. 1202 (American Chemical Society, 2015) pp. 17–43.
[14] M. M. Castellanos, N. J. Clark, M. C. Watson, S. Krueger, A. McAuley, and J. E. Curtis, The Journal of Physical Chemistry B **120**, 12511 (2016).
[15] Y. Zhai, N. S. Martys, W. L. George, J. E. Curtis, J. Nayem, Y. Z, and Y. Liu, Structural Dynamics **8**, 024102 (2021).
[16] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, and M. Feig, eLife **5**, e19274 (2016).
[17] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, in *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2014) pp. 41–53.
[18] M. M. Rickard, Y. Zhang, M. Gruebele, and T. V. Pogorelov, J. Phys. Chem. Lett. **10**, 5667 (2019).
[19] G. A. Voth, *Coarse-Graining of Condensed Phase and Biomolecular Systems* (CRC Press, 2008).
[20] J. K. Cheung and T. M. Truskett, Biophys. J. **89**, 2372 (2005).
[21] J. K. Cheung, P. S. Raverkar, and T. M. Truskett, J. Chem. Phys. **125**, 224903 (2006).
[22] V. K. Shen, J. K. Cheung, J. R. Errington, and T. M. Truskett, Biophys. J. **90**, 1949 (2006).
[23] J. K. Cheung, V. K. Shen, J. R. Errington, and T. M. Truskett, Biophys. J. **92**, 4316 (2007).
[24] V. K. Shen, J. K. Cheung, J. R. Errington, and T. M. Truskett, J. Biomech. Eng. **131**, 071002 (2009).
[25] V. Ramasubramani, T. Vo, J. A. Anderson, and S. C. Glotzer, J. Chem. Phys. **153**, 084106 (2020).
[26] M. A. Blanco, mAbs **14**, 2044744 (2022).
[27] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[28] A. Chaudhri, I. E. Zarraga, S. Yadav, T. W. Patapoff, S. J. Shire, and G. A. Voth, J. Phys. Chem. B **117**, 1269 (2013).
[29] M. A. Blanco, T. Perevozchikova, V. Martorana, M. Manno, and C. J. Roberts, J. Phys. Chem. B **118**, 5817 (2014).
[30] H. W. Hatch, J. Mittal, and V. K. Shen, J. Chem. Phys. **142**, 164901 (2015).
[31] H. W. Hatch, S.-Y. Yang, J. Mittal, and V. K. Shen, Soft Matter **12**, 4170 (2016).
[32] J. W. Wagner, J. F. Dama, A. E. P. Durumeric, and G. A. Voth, J. Chem. Phys. **145**, 044108 (2016).
[33] R. D. Mountain, H. W. Hatch, and V. K. Shen, Fluid Ph. Equilib **440**, 87 (2017).
[34] J. F. Dama, J. Jin, and G. A. Voth, J. Chem. Theory Comput. **13**, 1010 (2017).
[35] C. J. O'Brien, M. A. Blanco, J. A. Costanzo, M. Enterline, E. J. Fernandez, A. S. Robinson, and C. J. Roberts, Protein Eng. Des. Sel. **29**, 231 (2016).
[36] G. V. Barnett, M. Drenski, V. Razinkov, W. F. Reed, and C. J. Roberts, Anal. Biochem. **511**, 80 (2016).
[37] C. Calero-Rubio, A. Saluja, E. Sahin, and C. J. Roberts, J. Phys. Chem. B **123**, 5709 (2019).
[38] M. A. Woldeyes, C. Calero-Rubio, E. M. Furst, and C. J. Roberts, in *Protein Self-Assembly: Methods and Protocols*, Methods in Molecular Biology, edited by J. J. McManus (Springer New York, New York, NY, 2019) pp. 23–37.
[39] N. Skar-Gislinge, M. Ronti, T. Garting, C. Rischel, P. Schurtenberger, E. Zaccarelli, and A. Stradner, Mol. Pharmaceutics **16**, 2394 (2019).
[40] H. Shahfar, J. K. Forder, and C. J. Roberts, J. Phys. Chem. B **125**, 3574 (2021).
[41] S. Mahapatra, M. Polimeni, L. Gentiluomo, D. Roessner, W. Frieß, G. H. J. Peters, W. W. Streicher, M. Lund, and P. Harris, Mol. Pharmaceutics **19**, 508 (2022).
[42] J. K. Forder, A. J. Ilott, E. Sahin, and C. J. Roberts, AIChE J. **69**, e17965 (2023).
[43] M. Polimeni, E. Zaccarelli, A. Gulotta, M. Lund, A. Stradner, and P. Schurtenberger, APL Bioeng. **8**, 016111 (2024).
[44] J. K. Forder, V. Palakollu, S. Adhikari, M. A. Blanco, M. G. Derebe, H. M. Ferguson, S. A. Luthra, E. V. Munsell, and C. J. Roberts, Mol. Pharmaceutics **21**, 1321 (2024).
[45] A. Grünberger, P.-K. Lai, M. A. Blanco, and C. J. Roberts, J. Phys. Chem. B **117**, 763 (2013).
[46] M. A. Blanco, E. Sahin, A. S. Robinson, and C. J. Roberts, J. Phys. Chem. B **117**, 16013 (2013).
[47] C. Calero-Rubio, A. Saluja, and C. J. Roberts, J. Phys. Chem. B **120**, 6592 (2016).
[48] A. Chowdhury, J. A. Bollinger, B. J. Dear, J. K. Cheung, K. P. Johnston, and T. M. Truskett, Mol. Pharmaceutics **17**, 1748 (2020).
[49] J. J. Hung, W. F. Zeno, A. A. Chowdhury, B. J. Dear, K. Ramachandran, M. P. Nieto, T. Y. Shay, C. A. Karouta, C. C. Hayden, J. K. Cheung, T. M. Truskett, J. C. Stachowiak, and K. P. Johnston, Soft Matter **15**, 6660 (2019).
[50] B. J. Dear, J. A. Bollinger, A. Chowdhury, J. J. Hung, L. R. Wilks, C. A. Karouta, K. Ramachandran, T. Y. Shay, M. P. Ni-

eto, A. Sharma, J. K. Cheung, D. Nykypanchuk, P. D. Godfrin, K. P. Johnston, and T. M. Truskett, J. Phys. Chem. B **123**, 5274 (2019).
[51]N. Skar-Gislinge, F. Camerin, A. Stradner, E. Zaccarelli, and P. Schurtenberger, Molecular Pharmaceutics **20**, 2738 (2023).
[52]M. A. Blanco, H. W. Hatch, J. E. Curtis, and V. K. Shen, J. Pharm. Sci. **108**, 1663 (2019).
[53]P. Mereghetti, R. R. Gabdoulline, and R. C. Wade, Biophys. J. **99**, 3782 (2010).
[54]S. R. McGuffee and A. H. Elcock, PLoS Comput. Biol. **6**, e1000694 (2010).
[55]V. Prytkova, M. Heyden, D. Khago, J. A. Freites, C. T. Butts, R. W. Martin, and D. J. Tobias, J. Phys. Chem. B **120**, 8115 (2016).
[56]B. B. Majumdar, V. Prytkova, E. K. Wong, J. A. Freites, D. J. Tobias, and M. Heyden, J. Chem. Theory Comput. **15**, 1399 (2019).
[57]B. B. Majumdar, S. Ebbinghaus, and M. Heyden, J. Theor. Comput. Chem. **17**, 1840006 (2018).
[58]S. P. Carmichael and M. S. Shell, J. Chem. Phys. **139**, 164705 (2013).
[59]H. W. Hatch, W. P. Krekelberg, S. D. Hudson, and V. K. Shen, J. Chem. Phys. **144**, 194902 (2016).
[60]H. W. Hatch, N. A. Mahynski, R. P. Murphy, M. A. Blanco, and V. K. Shen, AIP Adv. **8**, 095210 (2018).
[61]R. P. Murphy, H. W. Hatch, N. A. Mahynski, V. K. Shen, and N. J. Wagner, Soft Matter **16**, 1279 (2020).
[62]H. W. Hatch and G. W. McCann, J. Res. Natl. Inst. Stand. Technol **124**, 1 (2019).
[63]I. A. Vakser, S. Grudinin, N. W. Jenkins, P. J. Kundrotas, and E. J. Deeds, Proc. Natl. Acad. Sci. U.S.A. **119**, e2210249119 (2022).
[64]S. Qin and H.-X. Zhou, J. Phys. Chem. B **123**, 8203 (2019).
[65]A. J. Schultz and D. A. Kofke, J. Chem. Phys. **157**, 190901 (2022).
[66]F. Platten, N. E. Valadez-Pérez, R. Castañeda-Priego, and S. U. Egelhaaf, J. Chem. Phys. **142**, 174905 (2015).
[67]A. Y. Xu, M. A. Blanco, M. M. Castellanos, C. W. Meuse, K. Mattison, I. Karageorgos, H. W. Hatch, V. K. Shen, and J. E. Curtis, J. Phys. Chem. B **127**, 8344 (2023).
[68]N. Sibanda, R. K. Shanmugam, and R. Curtis, Mol. Pharmaceutics **20**, 2662 (2023).
[69]M. G. Noro and D. Frenkel, J. Chem. Phys. **113**, 2941 (2000).
[70]H. W. Hatch, S. Jiao, N. A. Mahynski, M. A. Blanco, and V. K. Shen, J. Chem. Phys. **147**, 231102 (2017).
[71]N. A. Mahynski, S. Jiao, H. W. Hatch, M. A. Blanco, and V. K. Shen, J. Chem. Phys. **148**, 194105 (2018).
[72]J. I. Monroe, H. W. Hatch, N. A. Mahynski, M. S. Shell, and V. K. Shen, J. Chem. Phys. **153**, 144101 (2020).
[73]S. Yadav, T. M. Laue, D. S. Kalonia, S. N. Singh, and S. J. Shire, Mol. Pharmaceutics **9**, 791 (2012).
[74]E. Binabaji, J. Ma, and A. L. Zydney, Pharm. Res. **32**, 3102 (2015).
[75]J. K. Singh and D. A. Kofke, Phys. Rev. Lett **92**, 220601 (2004).
[76]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graph. **14**, 33 (1996).
[77]A. C. M. Young, J. C. Dewan, C. Nave, and R. F. Tilton, J. Appl. Crystallogr. **26**, 309 (1993).
[78]A. S. Parmar and M. Muschol, Biophys. J. **97**, 590 (2009).
[79]H. W. Hatch, N. A. Mahynski, and V. K. Shen, J. Res. Natl. Inst. Stan **123**, 123004 (2018).
[80]H. W. Hatch, D. W. Siderius, and V. K. Shen, J. Chem. Phys. (2024), in Press in the same special issue as this manuscript. Manuscript ID JCP24-SW-MCM2023-02594.
[81]R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, J. Comput. Chem. **28**, 1145 (2007).
[82]D. Sitkoff, K. A. Sharp, and B. Honig, J. Phys. Chem. **98**, 1978 (1994).
[83]T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, Nucleic Acids Res. **32**, W665 (2004).
[84]M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, J. Chem. Theory Comput. **7**, 525 (2011).
[85]F. Sciortino, S. Mossa, E. Zaccarelli, and P. Tartaglia, Phys. Rev. Lett **93**, 055701 (2004).
[86]Y. Liu and Y. Xi, Curr. Opin. Colloid & Inter. Sci. **39**, 123 (2019).
[87]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: the art of scientific computing* (Cambridge University Press, 2007).
[88]M. Lund, M. Trulsson, and B. Persson, Source Code Biol. Med. **3**, 1 (2008).
[89]K. M. Benjamin, J. K. Singh, A. J. Schultz, and D. A. Kofke, J. Phys. Chem. B **111**, 11463 (2007).
[90]F. J. Vesely, Journal of Computational Physics **47**, 291 (1982).
[91]E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, Proc. Natl. Acad. Sci. U.S.A. **89**, 2195 (1992).
[92]D. F. Rosenbaum and C. F. Zukoski, J. Cryst. Growth **169**, 752 (1996).
[93]H. Bajaj, V. K. Sharma, and D. S. Kalonia, Biophys. J. **87**, 4048 (2004).
[94]O. D. Velev, E. W. Kaler, and A. M. Lenhoff, Biophys. J. **75**, 2682 (1998).
[95]F. Bonneté, S. Finet, and A. Tardieu, J. Cryst. Growth **196**, 403 (1999).
[96]R. A. Curtis, C. Steinbrecher, M. Heinemann, H. W. Blanch, and J. M. Prausnitz, Biophys. Chem. **98**, 249 (2002).
[97]A. Rohatgi, "Webplotdigitizer: Version 4.7," (2024).
[98]R. A. Curtis, J. M. Prausnitz, and H. W. Blanch, Biotechnol. Bioeng. **57**, 11 (1998).
[99]W. Li, B. A. Persson, M. Lund, J. Bergenholtz, and M. Zackrisson Oskolkova, J. Phys. Chem. B **120**, 8953 (2016).
[100]W. Li, B. A. Persson, M. Morin, M. A. Behrens, M. Lund, and M. Zackrisson Oskolkova, J. Phys. Chem. B **119**, 503 (2015).
[101]B. A. Persson, M. Lund, J. Forsman, D. E. W. Chatterton, and T. Åkesson, Biophys. Chem. **151**, 187 (2010).
[102]W. Li, M. Morin, E. Gustafsson, B. A. Persson, M. Lund, and M. Z. Oskolkova, Soft Matter **12**, 9330 (2016).
[103]C. Bergonzo and D. T. Gallagher, J. Res. Natl. Inst. Stand. Technol **126**, 126012 (2021).
[104]D. T. Gallagher, C. V. Galvin, and I. Karageorgos, Acta Crystallogr. Sect. F Struct. Biol. **74**, 524 (2018).
[105]I. Karageorgos, E. S. Gallagher, C. Galvin, D. T. Gallagher, and J. W. Hudgens, Biologicals **50**, 27 (2017).
[106]D. R. Roe and C. Bergonzo, J. Comput. Chem. **43**, 930 (2022).
[107]J. M. Prien, H. Stöckmann, S. Albrecht, S. M. Martin, M. Varatta, M. Furtado, S. Hosselet, M. Wang, T. Formolo, P. M. Rudd, and J. E. Schiel, in *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 2. Biopharmaceutical Characterization: The NISTmAb Case Study*, ACS Symposium Series, Vol. 1201 (American Chemical Society, 2015) pp. 185–235, section: 4.
[108]D. A. Case, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York, and P. A. Kollman, *AMBER 2018* (University of California, San Francisco, 2018).
[109]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, J. Chem. Theory Comput. **11**, 3696 (2015).
[110]K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods, J. Comput. Chem. **29**, 622 (2008).

[111]H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[112]I. S. Joung and T. E. I. Cheatham, J. Phys. Chem. B **112**, 9020 (2008).
[113]D. R. Roe and B. R. Brooks, J. Chem. Phys. **153**, 054123 (2020).
[114]R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, J. Chem. Theory Comput. **9**, 3878 (2013).
[115]J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[116]C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, J. Chem. Theory Comput. **11**, 1864 (2015).
[117]T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[118]J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic, and B. O. Brandsdal, Chem. Phys. Lett. **384**, 288 (2004).
[119]R. J. Loncharich, B. R. Brooks, and R. W. Pastor, Biopolymers **32**, 523 (1992).
[120]D. J. Sindhikara, S. Kim, A. F. Voter, and A. E. Roitberg, J. Chem. Theory Comput. **5**, 1624 (2009).
[121]D. R. Roe and T. E. I. Cheatham, J. Chem. Theory Comput. **9**, 3084 (2013).
[122]J. I. Siepmann and D. Frenkel, Mol. Phys. **75**, 59 (1992).
[123]M. M. Castellanos, S. C. Howell, D. T. Gallagher, and J. E. Curtis, Anal. Bioanal. Chem. **410**, 2141 (2018).
[124]M. M. Castellanos, K. Mattison, S. Krueger, and J. E. Curtis, Anal. Bioanal. Chem. **410**, 2161 (2018).
[125]N. J. Clark, H. Zhang, S. Krueger, H. J. Lee, R. R. Ketchem, B. Kerwin, S. R. Kanapuram, M. J. Treuheit, A. McAuley, and J. E. Curtis, J. Phys. Chem. B **117**, 14029 (2013).
[126]K. R. S. Shaul, A. J. Schultz, and D. A. Kofke, J. Chem. Phys. **135**, 124101 (2011).
[127]A. Hoffmann and S. Grudinin, J. Chem. Theory Comput. **13**, 2123 (2017).
[128]H. Nguyen, D. R. Roe, and C. Simmerling, J. Chem. Theory Comput. **9**, 2020 (2013).