UNIVERSITY OF CALIFORNIA,
IRVINE

Multi-conformation Monte Carlo: a Method for Introducing Flexibility in Efficient Simulations of Many-protein Systems

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Chemical and Materials Physics

by

Vera D. Prytkova

Thesis Committee:
Professor Douglas Tobias, Chair
Associate Professor Rachel Martin
Professor Ioan Andricioaei

2017

TABLE OF CONTENTS

The work presented in this thesis is published in *J. Phys. Chem. B, 2016, 120(33), pp 8115-8126* by Vera Prytkova, Matthias Heyden, Domarin Khago, J. Alfredo Freites, Carter T. Butts, Rachel W. Martin, and Douglas J. Tobias

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THE THESIS

Multi-conformation Monte Carlo: a Method for Introducing Flexibility in Efficient Simulations of Many-protein Systems

By

Vera D. Prytkova

Master of Science in Chemical and Materials Physics

University of California, Irvine, 2017

Professor Douglas Tobias, Chair

We present a novel multi-conformation Monte Carlo simulation method that enables the modeling of protein-protein interactions and aggregation in crowded protein solutions. This approach is relevant to a molecular-scale description of realistic biological environments, including the cytoplasm and the extracellular matrix, that are characterized by high concentrations of biomolecular solutes (e.g., 300-400 mg/mL for proteins and nucleic acids in the cytoplasm of *Escherichia coli*). Simulation of such environments necessitates the inclusion of a large number of protein molecules. Therefore, computationally inexpensive methods, such as rigid-body Brownian dynamics or Monte Carlo simulations, can be particularly useful. However, as we demonstrate herein, the rigid-body representation typically employed in simulations of many-protein systems gives rise to certain artifacts in protein-protein interactions. Our approach allows us to incorporate molecular flexibility in Monte Carlo simulations at low computational cost, thereby eliminating ambiguities arising from structure selection in rigid-body simulations. We

benchmark and validate the methodology using simulations of hen egg white lysozyme in solution, a well-studied system for which extensive experimental data, including osmotic second virial coefficients, small-angle scattering structure factors, and multiple structures determined by x-ray and neutron crystallography and solution NMR, as well as rigid-body BD simulation results, are available for comparison.

INTRODUCTION

Biological environments, such as the cytoplasm and the extracellular matrix, are characterized by high concentrations of proteins and other biomacromolecular solutes (e.g., 300-400 mg/mL in the cytoplasm of *Escherichia coli*[1]). Under such crowded conditions, intermolecular interactions cannot be neglected and have a significant influence on the stability of folded proteins as well as their dynamics and aggregation propensities.[2]

To model protein-protein interactions and protein aggregation using computer simulations, multiple protein molecules must be included in the simulation system. Fully atomistic simulations including explicit representations of the aqueous solvent are currently only feasible for systems containing a limited number of biomolecular solutes (on the order of 10) and 100 ns timescales.[3] Brownian dynamics (BD) simulations employing an implicit representation of the solvent have emerged as a powerful approach to modeling many-protein systems on significantly longer timescales.[4-7] In BD simulations protein molecules are usually modeled as rigid bodies and their translational and rotational motions are generated with picosecond timesteps using the Ermak-McCammon algorithm.[8] Fast potential and force calculations are achieved through the use of pre-evaluated, constant potential terms on space-filling grids. This approach allows the simulation of solutions containing ~1,000 atomically detailed protein molecules for ~10-100 □s, which is long enough to obtain converged structural and thermodynamic properties for concentrated protein solutions in which the proteins are not aggregating strongly.[4-6] In addition to providing structural and thermodynamic

information, BD simulations have been used to investigate the effects of crowding on diffusion in protein solutions and a model of the *E. coli* cytoplasm.[5, 9]

Monte Carlo (MC) simulations are potentially an attractive alternative approach for the modeling of aggregating systems,[10, 11] or to generally improve the configurational sampling efficiency when the sampling of explicit dynamics is not the primary goal. MC simulations based on highly coarse-grained colloidal sphere protein models have been employed to investigate phase behavior in protein solutions and protein crystallization.[12-16] MC simulations of more detailed protein models with residue level coarse-graining have been used to study the effects of solution conditions and ion binding on protein-protein interactions,[17-19] as well as protein self-assembly.[20] As we will show below, Metropolis MC simulations[21] of atomically detailed proteins can be used to investigate the structural and thermodynamic properties of crowded protein solutions with at least as good sampling efficiency as BD simulations. For more efficient sampling of strongly aggregating systems, MC simulations offer the possibility of specialized trial moves designed to expedite the formation and destruction of clusters, such as in the aggregation-volume-bias MC method pioneered by Siepmann, Chen, and co-workers.[10, 11, 22]

Here, we evaluate the feasibility of performing MC simulations using the protein-protein interaction potential developed by Wade and co-workers[6] for BD simulations of many-protein systems. We use new experimental static light scattering (SLS) data for the optimization of the interaction potential parameters in simulations of solutions of hen egg white lysozyme (HEWL). The optimized

parameters were validated by comparing structure factors computed from the simulations to those derived from small-angle x-ray and neutron scattering measurements. In conventional rigid-body MC simulations using a single protein configuration, we find a strong dependence of both structural and thermodynamic properties on the specifics of the protein conformation. These results highlight the importance of incorporating protein conformational flexibility in the simulations. We have, therefore, implemented a new technique, which we refer to as multi-conformation Monte Carlo (mcMC); mcMC incorporates conformational flexibility by swapping protein conformations within a discrete library determined by clustering of protein configurations from an atomistic MD simulation of a single protein in explicit solvent. The approach is similar in spirit to the use of pre-evaluated libraries of molecular fragment conformations in configurational-bias Monte Carlo simulations.[23-26] However, in mcMC simulations sampling of intramolecular degrees of freedom is restricted to a set of discrete conformations, which allows the use of pre-evaluated potential grids for highly efficient energy calculations.[4-6] The HEWL solution simulations with mcMC show better agreement with experimental data compared to the results of scMC simulations using a single protein configuration, and eliminate the bias imposed by the use of a single structure.

CHAPTER 1. METHODS

Protein-protein interaction potential. The overall protein-protein interaction potential we employ, which was developed by Mereghetti et al. for many-protein Brownian dynamics simulations using the SDAMM software package by Mereghetti et al.,[6, 27] contains four contributions. The first two account for the interactions of the charges on one protein with the electrostatic potential of a second protein and an "electrostatic desolvation" penalty when the charges on one protein enter a low dielectric cavity of a second protein. The two electrostatic contributions contain an explicit dependence on the solution ionic strength. The third contribution is a short-ranged attractive "nonpolar desolvation" potential that mimics hydrophobic interactions, and the fourth term describes soft-core repulsive interactions between atoms on different proteins.

Prior to the simulations the potentials for each simulated protein conformation were pre-computed on cubic grids. To determine the appropriate grid sizes, we examined the convergence of the radial distribution functions, $g(r)$, of the protein centers-of-mass with grid size (Figure S1 in the Supporting Information) in single-conformation MC simulations (described below). The results reported herein were generated using 200 x 200 x 200 grids with a 1 Å spacing for all of the terms in the interaction potential; Figure S1 shows that these grids are sufficient to obtain converged $g(r)$s.

The electrostatic potential grids were computed at each ionic strength considered for an atomistic representation of the proteins with charges corresponding to the OPLS force field[28] by finite-difference solution of the non-

linear Poisson-Boltzmann equation using either the UHBD[29] or APBS[30] software packages. Dielectric constants of 78.4 and 2.0 for the solvent and protein, respectively, and ion exclusion radii of 1.5 Å were used. For increased computational efficiency, the number of charged protein sites involved in the evaluation of the electrostatic potential terms during the simulations was reduced by using the effective charge formalism of Gabdoulline and Wade[31].

We used the parameters in the potential function reported by Mereghetti et al.[6] with the following exceptions: (1) the empirical scaling coefficients of the electrostatic and nonpolar desolvation potentials were varied and optimized by comparing simulated osmotic virial coefficients with experimental data (see Figure S2). Unless otherwise noted, the default scaling parameters of 0.36 (unitless) and −0.0090 kcal/mol/Å^2 were used for the electrostatic and nonpolar desolvation terms, respectively; (2) we increased the parameter σ in the soft-core repulsion potential from 3 Å to 10 Å in order to increase the energetic penalty of overlapping protein atoms. This was necessary because MC trial moves with such unfavorable configurations need to be rejected. With the original parameter, the energetic penalty for overlapping protein molecules was too small, allowing compensation via electrostatic terms. In BD simulations, such configurations are not accessible, allowing a lower penalty and a smoother soft-core repulsion that, in turn, prevents the occurrence of large forces.

Single-conformation Monte Carlo simulations. We implemented single-conformation MC (scMC) simulations based on translational and rotational trial moves of randomly selected molecules in the SDAMM software package for BD

simulations of many protein systems.[6] The step size of the translational and rotational trial moves adapts during the simulation to yield an acceptance ratio of roughly 50% to produce efficient sampling under all conditions. Trial moves are accepted with a probability given by the Metropolis criterion:[21]

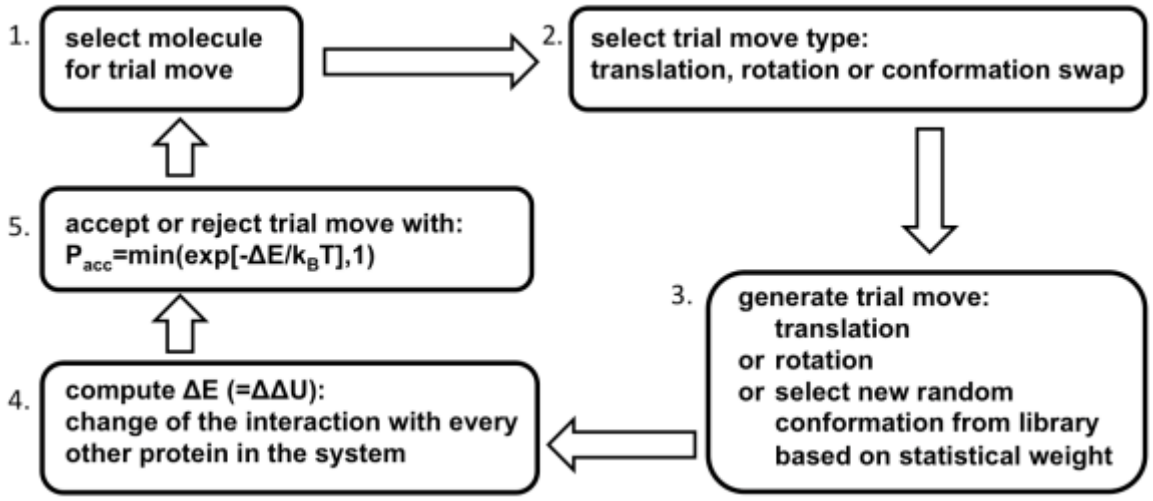
$$P_{acc} = \min\left[1,\ \exp\left(\frac{-\Delta U}{k_B T}\right)\right], \quad (1)$$

where ΔU is the difference in the protein-protein interaction potential between the current and trial configurations, k_B is Boltzmann's constant, and T is the temperature.

We performed scMC simulations of HEWL solutions over a range of protein concentrations and ionic strengths (Table S1) using three different structures obtained from neutron diffraction[32, 33] (PDB IDs 1IO5 and 1LZN) as well as solution NMR[34] (PDB ID 1E8L, model 1) experiments, all of which contain information on protonation states and proton coordinates in addition to the heavy atom coordinates. 1LZN has two protonated glutamates and a total charge of +11e, while 1IO5 and 1E8L carry a total charge of +9e.

Multi-conformation Monte Carlo simulations. To incorporate flexibility of the simulated proteins, we introduce a Monte Carlo trial move that, in addition to translational and rotational trial moves, attempts to swap protein conformations drawn from within a discrete library of conformations. This library is generated by clustering of protein conformations from an atomistic MD simulation of a single protein in explicit solvent. The candidate conformation is selected at random from the library with probability proportional to the population size of the corresponding

cluster in the MD simulation (see below). As in the scMC simulations, a standard Metropolis acceptance criterion is used. This approach ensures that, in the dilute-solution limit (i.e., no protein-protein interactions), the resulting distribution of conformational states converges to the distribution observed in the MD simulation of a single solvated protein. In the case of interacting proteins, the distribution of protein conformations can change due to the stabilization/destabilization of conformations in bound states. Potential grids are pre-computed for each of the structures in the library. The size of the library is limited only by the memory requirements of these grids. The conformational swap moves allow sampling of the most favourable conformations of interacting proteins at a computational cost that is the same as that of the rigid-body translational or rotational moves. The algorithm is sketched in Scheme 1.



Scheme 1. Sketch of the multi-conformational Monte Carlo algorithm for the simulation of flexible proteins using a library of conformations and standard translational and rotational trial moves.

Various approaches could be employed to generate a suitable library of conformations. Here, we used a 150 ns atomistic MD simulation trajectory of a single protein in explicit solvent using a HEWL solution NMR structure[34] (PDB ID 1E8L, model 1) as the initial configuration. To generate the library of protein conformations, we computed the heavy (not hydrogen) atom root-mean squared deviation (RMSD) matrix between configurations saved every 10 ps, and employed a simple clustering algorithm[35] using a 1 Å RMSD cutoff to extract the 50 most populated clusters. By including all non-hydrogen atoms, we ensure that backbone and side chain fluctuations give rise to distinct conformations in the library. The cluster centroids constitute the library of conformations, and the cluster populations were used to assign a statistical weight to each conformation in the library.

The MD trajectory was generated using the GROMACS software package.[36] The OPLS all-atom force field[28] was used for the protein and ions, and the TIP3P model[37] was used for water. The charge of the protein was neutralized with chloride ions, the system was solvated by 3579 water molecules, and periodic boundary conditions were applied in three dimensions. Short-ranged interactions were truncated with a 9 Å cutoff, while long-ranged electrostatic interactions were computed with the smooth particle-mesh Ewald method[38] on a 1.2 Å real-space grid. Covalent bonds and the geometry of water molecules were constrained with the LINCS[39] and SETTLE[40] algorithms, respectively. After an initial energy minimization, the system was equilibrated for 100 ps using a 1 fs integration time step in the isothermal-isobaric ensemble with harmonic restraints on protein heavy atom

positions using a force constant of 1000 kJ/mol/nm^2, followed by a 1 ns unrestrained equilibration using a 2 fs time step. During equilibration, a Berendsen[41] weak coupling thermostat and barostat was employed with time constants of 0.5 ps and 1.0 ps, respectively, and 300 K and 1 bar target values. The production simulation of 150 ns duration was generated with a 2 fs time step and the Nosé-Hoover thermostat[42] for temperature control and the Parrinello-Rahman barostat[43] for pressure control.

Calculation of osmotic second virial coefficients and structure factors. The osmotic second virial coefficient, B_2, is the second-order coefficient in the Taylor series expansion of the osmotic pressure in terms of number density;[44] it provides one of the very few experimental measures of pairwise interactions between protein molecules in solution: $B_2 < 0$ implies attractive interactions, while $B_2 > 0$ implies repulsive interactions, and the magnitude of B_2 quantifies the strength of the interactions. The osmotic second virial coefficient is computed from simulations according to:[44]

$$B_2 = -2\pi \int_0^\infty \left[g(r) - 1 \right] r^2 \, dr , \qquad (1)$$

where $g(r)$ is the radial distribution function of the protein centers-of-mass. Here, the radial distribution function describes a potential of mean force between two molecules, $W(r) = -k_B T \ln g(r)$, which includes averaging over all possible orientations and conformations of both molecules.

The structure factor, $S(q)$, is an interference function that arises from interparticle interactions and can be extracted from small-angle x-ray and neutron

scattering measurements. The structure factor is also readily computed from the protein-protein radial distribution function according to:[45]

$$S(q) = 1 + 4\pi\rho \int_0^\infty \left[g(r) - 1 \right] \frac{\sin(qr)}{qr} r^2 \, dr , \qquad (2)$$

where q is the modulus momentum transfer vector, and ρ is the solution density.

Sample preparation and static light scattering experiments. Lyophilized hen egg white lysozyme (Cat. No. 195303) was purchased from MP Biomedicals (Solon, OH). Lysozyme was dissolved in 10 mM sodium phosphate buffers containing 0.05% sodium azide (pH 4.7 and 6.9) with NaCl concentrations of 50, 75, 100, 125, 150, 200, 250, and 300 mM for a final protein concentration of 50 mg/mL. Serial dilutions were performed to prepare samples with protein concentrations ranging from 2.5 to 50 mg/mL for light scattering measurements. The concentrations were checked by UV absorbance measurements using ε = 2.64 mL mM^{-1} cm^{-1} at 280 nm. A Dawn HELEOS multi-angle light scattering instrument and an Opitlab rEX refractive index detector (Wyatt Technology, Santa Barbara, CA) were used to collect the data required for experimental B_2 determination. Samples were injected using the batch-mode technique from lowest to highest concentrations after filtering to ensure monodispersity.

Estimation of osmotic second virial coefficients from static light scattering data. Scattering intensity data at each concentration was processed to remove artifacts caused by sample injection, and the median of the remaining observations employed as the scattering intensity measurement for each detector. Medians were also taken for the refractive index increment at each concentration;

readings at concentrations greater than 0.02 g/mL exceeded the range of the differential refractometer, and were treated as missing for purposes of analysis.

For small particles in dilute solution, Zimm's[46] second order expansion of light scattering intensity (in terms of the excess Rayleigh ratio, R_q) with respect to concentration leads to the approximation:[47]

$$\frac{Kc}{R_q} \approx \frac{1}{MP(q)} + 2A_2 c, \qquad (4)$$

where $K = 4\pi^2 \left(dn/dc\right)^2 n_0^2 / N_A \lambda_0^4$, with dn/dc the refractive index increment, n_0 the solvent refractive index, λ_0 the vacuum wavelength of the incident light, N_A Avogadro's number, M the (mass weighted) mean particle mass, c the protein concentration, $P(q)$ a size-specific factor that depends upon the detection angle θ relative to the angle of incidence, and $A_2 = B_2 N_A / M^2$ is the osmotic second virial coefficient in a power series expansion of the osmotic pressure in terms of concentration. The $P(q) \to 1$ limit is realized as the particle radius of gyration r_g approaches 0; for monomeric or small oligomeric particles with $r_g << \lambda_0$, angular dependence is negligible, and we employ this limit here. Note that no angular dependence was detected in our experiments, which is consistent with predictions for a beam wavelength of 658 nm and $r_g \approx 0.14$ nm.[48]

As B_2 represents a very small deviation in local effective particle density (relative to uniform mixing), it is challenging to estimate with high precision. We employ a number of techniques to address this issue. Given multiple observations of R_q at varying concentration, it is natural to estimate A_2 by regression of Kc/R_q on

2c; when the scattering particles are monodisperse and of known mass, improved precision can be obtained by employing $Kc/R_q - 1/M$ as the response and fitting a zero-intercept model. When particles are known to be monodisperse but the oligomer size is not known, greater precision can still be obtained by fitting models to k-mers of orders 1, 2, … and selecting the k that minimizes the squared error in the predicted scattering intensity. As our samples were filtered to ensure monodispersity (with verification by dynamic light scattering) and the monomer mass is known, we employ this strategy here. This estimate also depends upon dn/dc, which must itself be estimated by regressing n_c (the measured refractive index at concentration c) against c. Because n_c is in practice far more reliably measured than c itself, further gains in precision can be obtained by using the refractive index data to correct the measured concentration values prior to estimation of A_2 (i.e., regressing c on n_c and employing the predicted $\hat{c}$ values in place of c). Combining the above leads to the following multi-stage procedure for estimating A_2: (1) regress n_c on c to obtain an estimate of dn/dc ($\widehat{dn/dc}$); (2) regress c on n_c to obtain corrected concentration estimates $\hat{c}$; (3) for $k \in 1, 2, \ldots$, regress $Kc/R_q - 1/(kM)$ on $2\hat{c}$ to obtain $\widehat{A_2}|k$, selecting the k leading to the minimum squared error in R_q and associated $\widehat{A_2}$ for the final B_2 estimate.

Classical estimates of the precision of $\widehat{A_2}$ are problematic both because of the interdependence of $\widehat{A_2}$ on $\widehat{dn/dc}$ and because of the contribution of concentration to both sides of the regression. Here, we instead employ a non-parametric bootstrap procedure (using the boot library for R[49]) to estimate confidence intervals.

Specifically, the above procedure was repeated for 5000 random with-replacement joint resamples of the refractive index and scattering intensity data (with sample sizes preserved for each subset), and 95% confidence intervals were estimated from the resulting bootstrap replicates using the bias-corrected/adjusted percentile (BCa) method of Efron.[50] These are shown in Figures 3 and 10 as vertical lines. Oligomer size estimates were further inspected by examination of bootstrap standard errors for signs of instability; samples were estimated to be monomeric for all replicates in all conditions examined.

CHAPTER2. RESULTS AND DISCUSSION

Single-configuration Monte Carlo simulations. In protein solution simulations at low to medium concentrations (e.g., < 100 mg/mL), translational MC trial move steps can be significantly larger than BD time steps, increasing the efficiency of configurational sampling. This is illustrated in Figure 1, which displays the convergence of the radial distribution function between protein centers-of-mass in scMC and BD simulations of HEWL under identical conditions at a concentration of 10 mg/mL. The sampling efficiency in the scMC simulations is increased by roughly two orders of magnitude compared to BD, producing a converged radial distribution function after 100k MC cycles (one cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation).

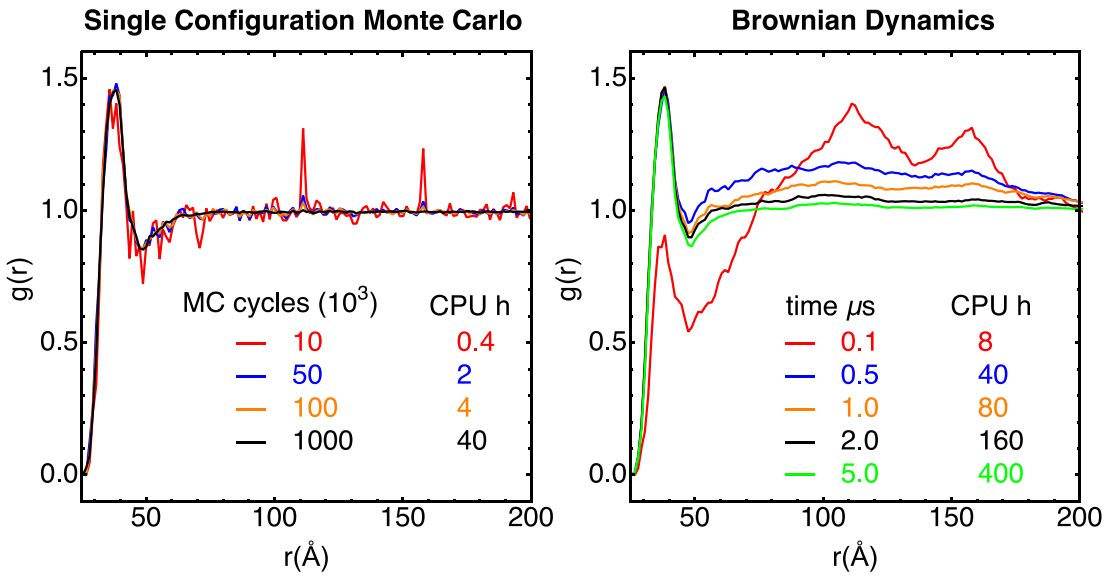


Figure 1. Convergence of the protein-protein radial distribution function from scMC (left panel) and BD (right panel) simulations of HEWL solutions containing 200 rigid proteins (1LZN structure) at a concentration of 10 mg/mL and ionic strength of 200 mM. The

sampling efficiency, in terms of CPU hours needed to generate a converged radial distribution function, is roughly two orders of magnitude higher for scMC than BD simulations (CPU h on single Intel Xeon E5430 processor with 2.66 GHz). scMC simulation lengths are expressed in MC cycles; a single MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

The scMC sampling efficiency advantage over BD vanishes at high concentrations, when the step size of translational MC trial moves need to be reduced in order to maintain acceptance ratios on the order of 50%, as shown in Figure 2 for a 169 mg/mL HEWL solution. However, we point out that MC simulations offer additional advantages for the simulation of slowly converging systems, e.g., in the event of protein aggregation. For example, biased sampling schemes, such as the aggregation volume biased MC technique developed by Siepmann and co-workers,[10, 11] can be readily implemented to improve the sampling of the formation and destruction of clusters.

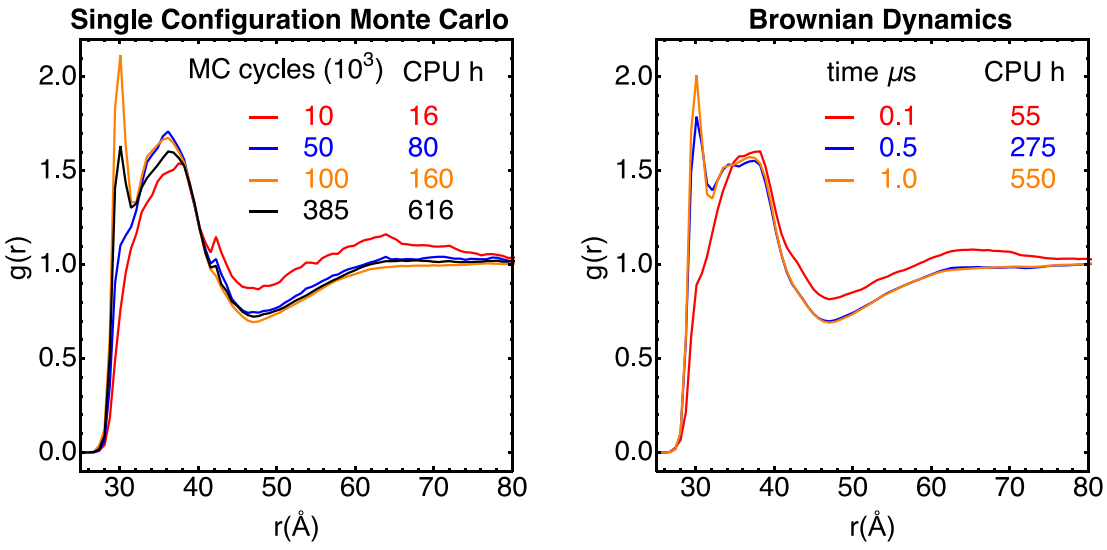


Figure 2. Convergence of the protein-protein radial distribution function from scMC (left panel) and BD (right panel) simulations of HEWL solutions containing 475 rigid proteins (1E8L structure) at a concentration of 169 mg/mL and ionic strength of 100 mM. The sampling efficiency is similar for both scMC and BD simulations at this concentration (CPU h on single Intel Xeon E5430 processor with 2.66 GHz); scMC simulation lengths are expressed in MC cycles; a single MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Previous BD simulations of HEWL solutions using the same[6] or similar[4] protein interaction potentials were validated using osmotic second virial coefficients as a function of solution ionic strength reported in the literature.[45] Here, we report two new sets of B_2 values from SLS measurements on HEWL at two different pH values (see Figure 3). Our experimental results are in good agreement with the literature[45, 51-54] and are consistent with Derjaguin-Landau-Verwey-Overbeek (DLVO) theory for colloidal systems.[52, 55]

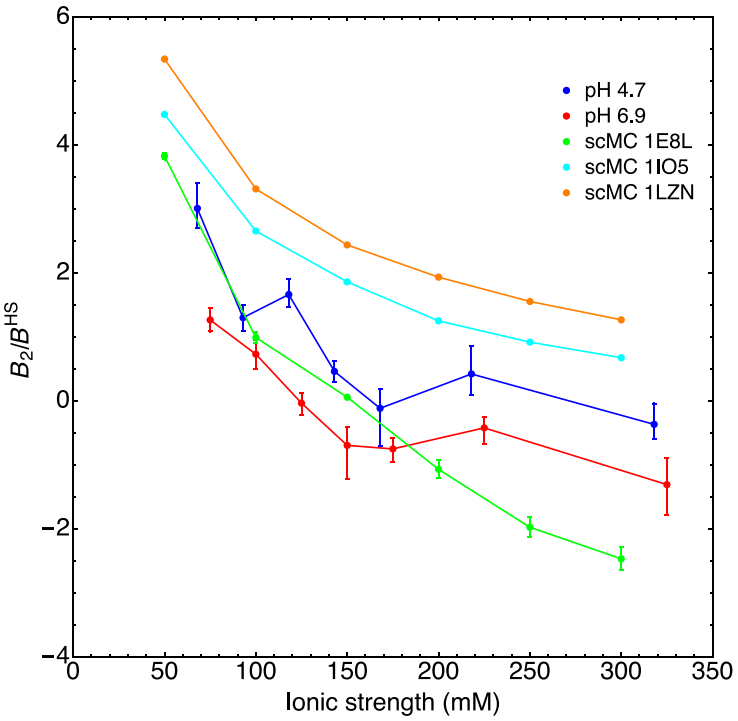


Figure 3. HEWL second virial coefficient (B_2; in units of B_2 for hard spheres, B^{HS} = 4 x protein volume) estimated from static light scattering experiments and scMC simulations as a function of solution ionic strength. Solution pH values of 4.7 and 6.9 reported in the experimental measurements correspond to estimated HEWL net charges of +10e and +8e, respectively.[56] Proteins in solution simulations based on crystal structures (1IO5 and 1LZN, net charge +9e and +11e, respectively) are overall more repulsive than indicated by experiments. HEWL in simulations based on the solution NMR structure (1E8L), which has the same net charge as 1IO5, are overall more attractive than in crystal structure simulations, and appear to be consistent with experiments (see text for more details) at ionic strength values less than ~0.2 M. However, at higher ionic strengths, the proteins appear to be more attractive than indicated by experiments.

Comparison of the ionic strength dependence of the B_2 values computed from scMC simulations at a concentration of 10 mg/mL using different input structures

reveals that the protein-protein interactions are strongly dependent on the choice of structure, and shows that a single structure is not able to reproduce the trend in the experimental data (Figure 3). Electrostatic repulsion is overestimated and, hence, B_2 is too large, in the two simulations based on crystal structures (1IO5 and 1LZN, net charge +9e and +11e, respectively). At ionic strength values below ~0.2 M, the B_2 values obtained from the solution NMR structure (1E8L, net charge +9e) follow the correct qualitative trend compared to the experimental data, but the preponderance of electrostatic interactions is evident by the lack of a plateau at ionic strength values above ~0.2 M. Notably, the protein-protein interactions are significantly more attractive in the 1E8L simulations than in the simulations based on the crystal structure with the same net charge (1IO5).

In addition to the overall increased protein-protein attractive interaction, the corresponding radial distribution functions between protein centers exhibit a spurious peak at $r \sim 30$ Å in simulations of the 1E8L HEWL structure, corresponding to specific protein-protein contacts that are not present in the simulation based on the 1IO5 and 1LZN crystal structures (Figure 4). Furthermore, the differences between the 1E8L and 1IO5 radial distribution functions are much greater than the differences between the 1IO5 and 1LZN simulations, suggesting that the specifics of protein conformations can have a much more dramatic effect than the total charge.

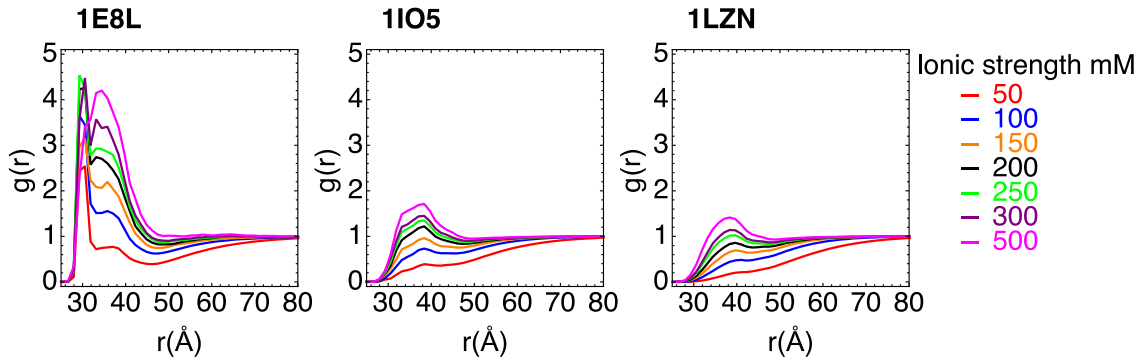


Figure 4. Protein-protein radial distribution functions $g(r)$ from the scMC simulations from which the B_2 values plotted in Figure 3 were obtained (10 mg/mL). In addition to a spurious peak at ~30 Å in simulations based on the 1E8L structure, there is more contrast in the protein-protein interactions when comparing the provenance of the protein structure (e.g., 1E8L vs. 1IO5) than when comparing different protein net charges (1IO5 vs. 1LZN).

In Figure 5, we compare isosurfaces of spatial distribution functions (SDFs), specifically, distributions of the density of other protein centers-of-mass around a tagged central protein, normalized by the bulk density at a protein concentration of 10 mg/mL with an ionic strength of 100 mM. Despite the apparent similarity of the radial distribution functions for scMC simulations based on the two crystal structures 1IO5 and 1LZN, the SDF from the simulation based on the 1IO5 structure reveals interaction sites that are not observed in the SDF from the simulation based on the more repulsive 1LZN structure. Similarly, the comparison to the SDF obtained from the simulation based on the solution NMR structure (1E8L) shows not only an overall increased attraction, but also additional locations of preferred contact sites.

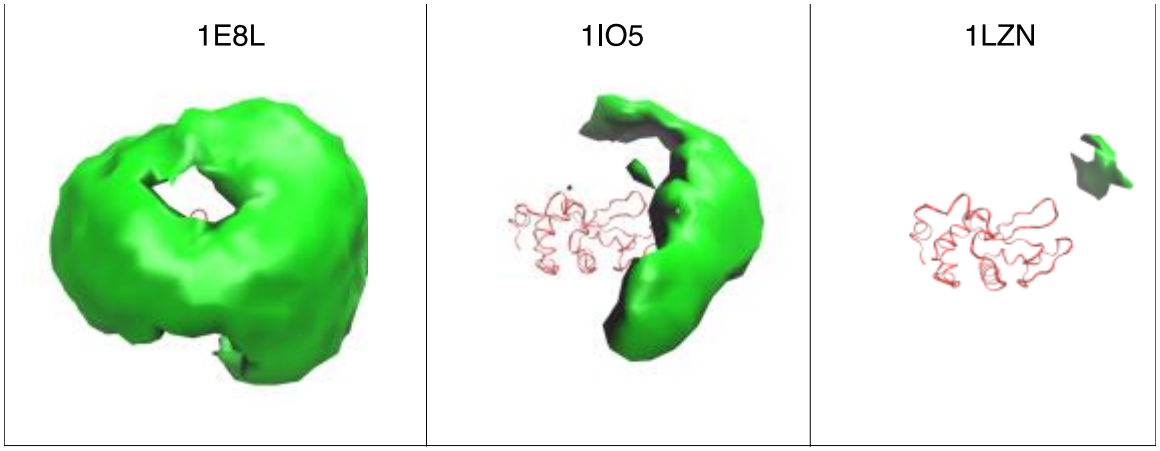


Figure 5. Spatial distribution functions (see text for details) for scMC simulations at 100 mM ionic strength. The green surfaces represent regions of 1.7 times the bulk density computed on a 160 Å cubic grid with a 4 Å grid spacing centered on and aligned with respect to the reference molecule shown in red. In addition to the greater protein-protein attraction in the 1E8L simulation, the simulations based on the two crystal structures exhibit interaction sites that are distinct from each other.

The SDF for the 1E8L simulation as shown in Figure 5 computed with an increased resolution on a 1 Å grid (Figure 6A) reveals two highly localized sites with a more than 1500-fold increase of the concentration relative to the bulk solution (corresponding to a stabilization in free energy of –4.3 kcal/mol). The corresponding dimer structures, extracted from the simulation trajectories, show binding motifs characterized by energetically favorable contacts of charged side chains (Figure 6B). Such specific binding motifs were not observed in simulations of the 1LZN and 1IO5 HEWL structures. The highest local density represents only 9-fold and 20-fold increases relative to the bulk density in scMC simulations based on the 1LZN and 1IO5 structures, respectively.

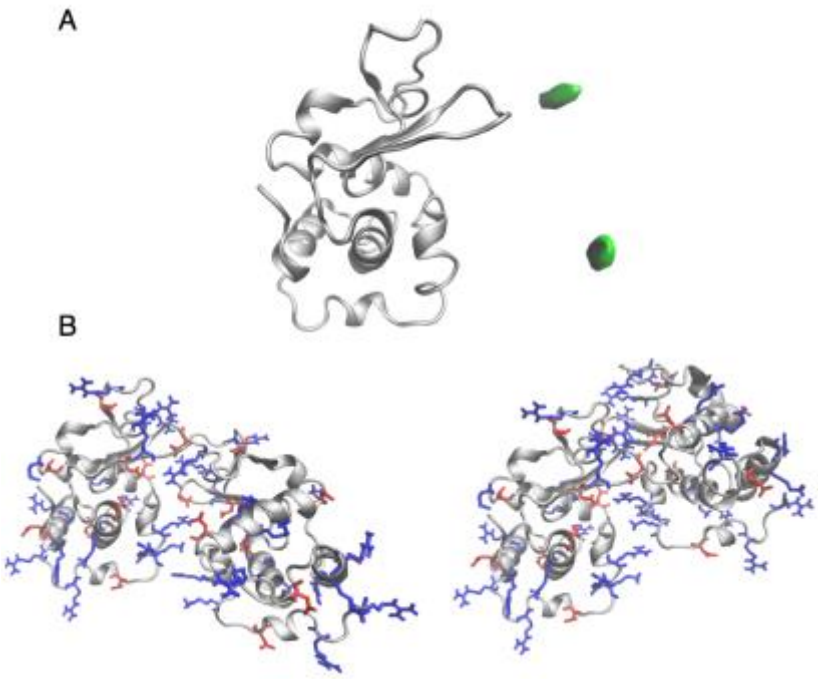


Figure 6. (A) Spatial distribution function for the 1E8L simulation at 100 mM ionic strength computed on a 100 Å cubic grid with a grid spacing of 1 Å. The green isosurfaces correspond to a 1500-fold increase of the local density relative to the bulk density. (B) Dimer configurations corresponding to the regions of increased concentration shown in (A). Solvent exposed basic (blue) and acidic (red) side chains are rendered in ball-and-stick representation.

The fact that 1IO5 and 1E8L structures have the same total charge and protonation state indicates that the highly specific binding observed in the 1E8L simulations originates in the specific arrangements of side chains on the protein surface, which is a consequence of the differences in structure determination methods. The 1E8L structure represents a protein in an aqueous solution environment where the charged side chains are in more extended conformations than those in the crystal environment required for the determination of the 1IO5

structure via neutron diffraction. A superposition of the 1IO5 and 1E8L structures (Figure 7) indicates, apart from the almost perfect alignment of the protein backbone, that the basic side chains tend to be significantly more extended and solvent exposed in the solution NMR structure than in the crystal, where they are, on average, more folded onto the protein surface. The solvent exposed basic side chains will be floppy and explore multiple conformations in solution. However, in scMC simulations as well as in rigid-body BD simulations, they are rigid, effectively losing their conformational entropy. This leads to enhanced favorable inter-protein interactions between exposed side chains of opposite charge, as observed in Figure 6, which in turn create highly specific binding motifs. This behavior is unrealistic, as the conformational entropy of the solvent-exposed side chains should decrease the population of such highly specific conformations. Thus, our results demonstrate that artifacts in inter-protein interactions may arise due to including only single side chain conformations in rigid-body BD or MC simulations of many-protein systems.

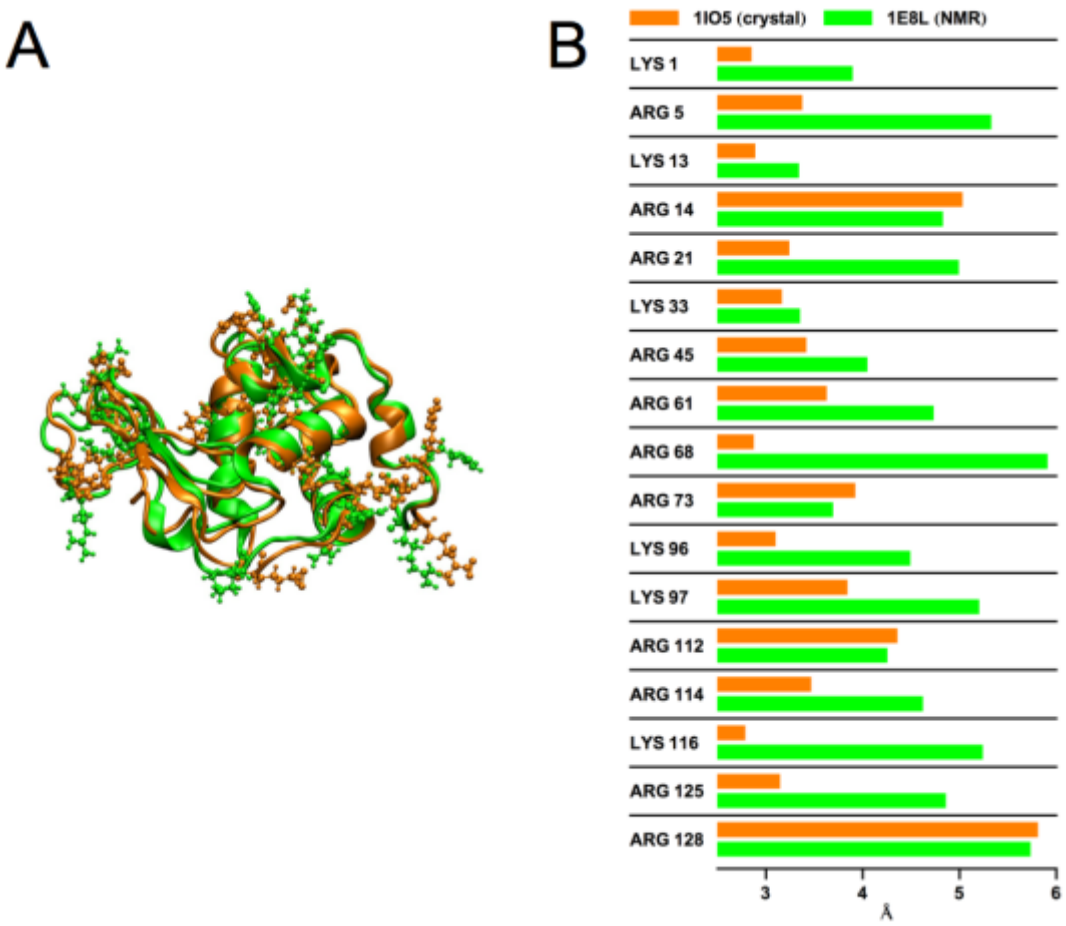


Figure 7. Conformations of basic side chains in the neutron diffraction structure 1IO5 (orange) and the solution NMR structure 1E8L (green) of HEWL (A), and distances between the centers of charge and the protein surface (B), indicate that they are more extended and solvent-exposed in the solution NMR structure than in the crystal structure. The behavior of these side chains in solution is poorly modeled with a single protein configuration.

Multi-conformation Monte Carlo simulations. To eliminate the dependence of the simulation results on the specifics of the protein conformation, we implemented the mcMC method, which introduces conformational flexibility in the protein molecules by allowing them to convert from one conformation to another within a structural library generated from an all-atom MD simulation (see

Methods for details). While the degree of flexibility introduced this way is obviously limited and not able to describe the entire protein conformational space, we posit that any reasonable choice of the ensemble will provide a significant improvement over the modeling of proteins in solution as single-conformation rigid bodies.

Our library consists of 50 structures that collectively represent a broad sampling of the basic and acidic side chain conformations without significant structural changes to the HEWL backbone (backbone RMSD < 2 Å). The variation of the positions of the positive and negative charge centers within the ensemble is shown in Figure 8A. The statistical weights associated with the library of conformations, obtained from the relative population of each cluster in the MD trajectory and used in the generation of conformational swap trial moves, are depicted in Figure 8B.

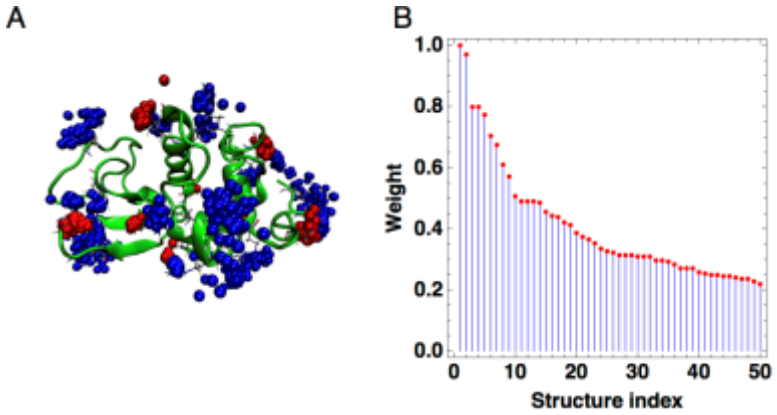


Figure 8. Structural library of HEWL employed in mcMC simulations obtained from a 150 ns all-atom MD simulation trajectory. (A) Superposition of the charge centers of basic (blue) and acidic (red) amino acid side chains in the 50 conformations that comprise the library. B) The statistical weights of individual structures, determined by their relative populations along the MD trajectory, are used as probabilities for the generation of conformational swap trial moves in the mcMC simulations.

To maximize agreement between experimental second virial coefficients and simulations with the mcMC algorithm, we adapted the protein-protein interaction potential by rescaling the empirical nonpolar desolvation (ND) potential term (as also described in other studies with the employed interaction potential[6, 27]). This term describes a short-ranged, uniform attraction between protein surfaces and mimics hydrophobic interactions. We also considered the effects of simultaneous rescaling of the empirical electrostatic desolvation, ED (repulsive, non-uniform), and nonpolar desolvation, ND (attractive, uniform), terms. We found that the two desolvation terms have compensating effects, resulting in a set of optimal ED and ND scaling factor pairs that include the default scaling value of the ED term (Figure S2). Therefore, we opted for leaving the ED scaling factor at its default value and explored in more detail the consequences of variations in the ND scaling factor.

Figure 9 shows the effects of changing the strength of the nonpolar desolvation potential on the protein-protein radial distribution function. When the default value of the scaling parameter (ND = −0.0090 kcal/mol/Å^2) is used, the resulting radial distribution functions (Figure 9, left panels) are very similar to the ones obtained from scMC simulations based on the 1IO5 structure under identical conditions (middle panels of Figures 4 and 5), suggesting a comparable radially averaged interaction potential between the HEWL proteins in both simulations. However, the SDFs at 100 mM ionic strength again show significant differences between the two simulations (compare the left panel of Figure 9B with the middle panel of Figure 5), indicating the importance of side chain conformation and flexibility in determining the preferred binding geometries. The radial distribution

functions from mcMC simulations with increased strengths of nonpolar desolvation interactions (ND = −0.0098 kcal/mol/Å^2 and ND = −0.0100 kcal/mol/Å^2) and the 1E8L scMC simulation (Figure 4, left panel) exhibit comparable main peaks at ~35 Å separation distance. However, the spurious peak at ~30 Å is absent in the mcMC simulations, as expected for simulations with flexible side chains.

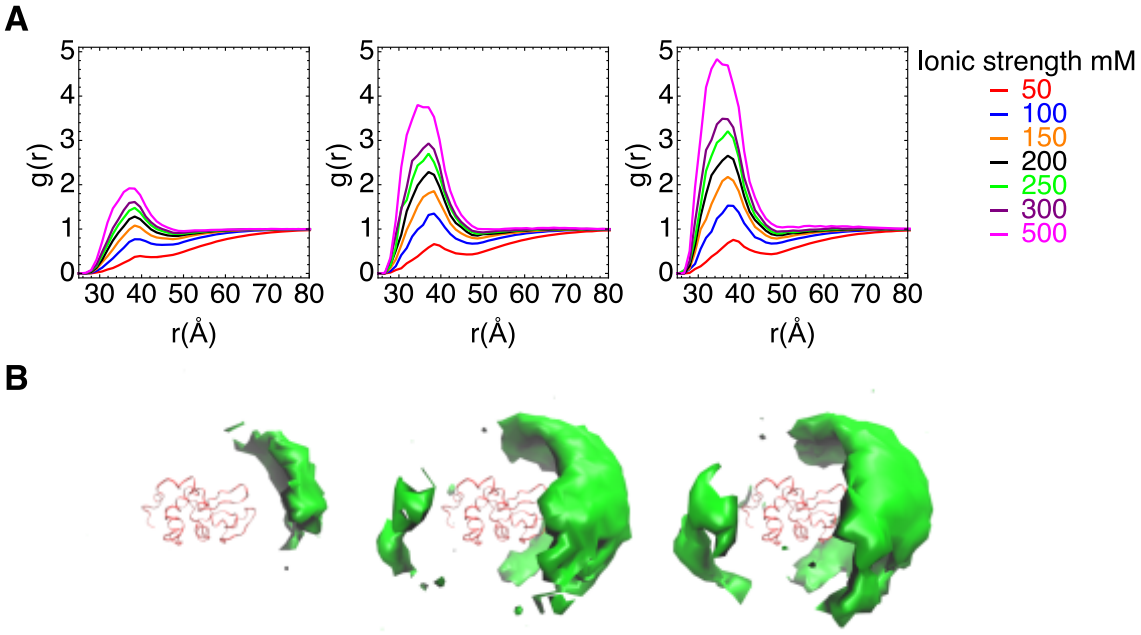


Figure 9. (A) Protein-protein radial distribution functions from HEWL mcMC simulations at three different values of the scaling parameter of the nonpolar desolvation potential, ND (from left to right, ND = −0.0090 kcal/mol/Å^2, −0.0098 kcal/mol/Å^2, −0.0100 kcal/mol/Å^2). (B) Isosurface contours (green) at 1.7 times the bulk density in the corresponding spatial distribution functions at 100 mM ionic strength.

The experimental osmotic second virial coefficients are reproduced well by the mcMC simulations with stronger nonpolar desolvation interactions (Figure 10). Notably, however, the SDFs, which define the preferred protein-protein binding interfaces, apart from the overall interaction strength, do not depend on the scaling factor of the nonpolar desolvation term (Figure 9B) as expected given the uniform

attraction relative to the solvent accessible surface area described by the scaled ND potential term. In the scMC simulation that employed the 1E8L solution NMR structure, the agreement with the experimental B_2 values at ionic strength below 200 mM was fortuitously achieved via increased exposure of the polar and charged side chains with zero conformational entropy. The exposure of polar and charged side chains is comparable for the structures in the ensemble used for the flexible mcMC simulations, as they are obtained from MD simulations in an explicit solvent environment. Thus, in contrast to the scMC case, in the mcMC simulation energetically highly favorable protein-protein interactions between individual structures lead to a compensating decrease of the conformational entropy, thus weakening the total binding affinity.

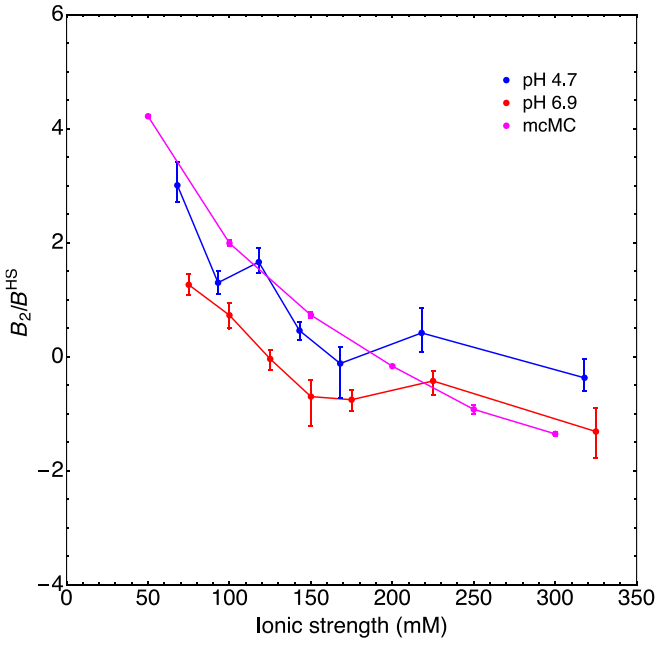


Figure 10. HEWL second virial coefficient (B_2; in units of B_2 for hard spheres, B^{HS} = 4 x protein volume) estimated from static light scattering experiments and mcMC simulations as a function of solution ionic strength at a protein concentration of 10 mg/mL. The mcMC

simulations were performed using the optimal value of the nonpolar desolvation strength determined from the experimental estimates (ND = −0.0098 kcal/mol/Å^2; see Supporting Information for more detail on the potential parameter optimization).

While osmotic second virial coefficients are determined at low concentrations (i.e., 2.5–50 mg/mL), structure factors allow us to validate our simulations against experimental data obtained at high protein concentration. Structure factors are interference functions that arise from protein-protein interactions in small-angle x-ray and neutron scattering measurements on protein solutions; peaks in the structure factors occur at values of the wave-vector transfer, $q \sim 2\pi/d$, corresponding to preferred interactions on length scales d. Results from mcMC simulations with variable scaling of the nonpolar desolvation potentials are shown in Figure 11 in comparison to experimental results[57, 58] and a previous BD simulation study by McGuffee *et al*.[4] The concentration used in all cases is 169 mg/mL. In order to compare our simulations with experiments at low ionic strength, we employed an ionic strength of 50 mM to ensure a sufficient decay of the electrostatic interactions within the employed potential grids. We note that the experimental studies have been carried out at a neutral pH with a slightly different protonation state (charge of +8e), while the HEWL proteins in our mcMC simulations carry a charge of +9e. Thus, minor differences between the simulation and experimental data are expected based on slight differences in solution conditions.

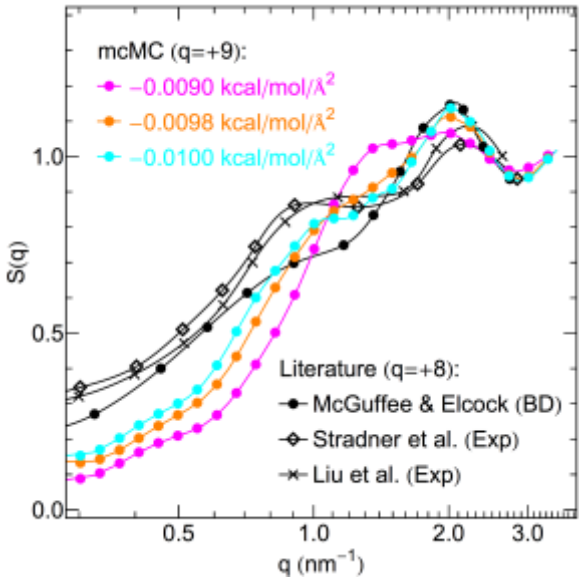


Figure 11. Structure factors from mcMC simulations (with varying nonpolar desolvation strength) of a HEWL solution at 169 mg/mL and 50 mM ionic strength in comparison to experimental data and a previous BD simulation, both at neutral pH. The choice made for the nonpolar desolvation strength parameter after optimization using B_2 estimates from low-concentration simulations (ND = −0.0098 kcal/mol/Å²) also produces the best match to experimental structure factors at high concentration.

The main peak at q = 2.0 nm^{-1} and the shoulder at q = 0.9–1.0 nm^{-1} are qualitatively reproduced by the mcMC simulations with modified nonpolar desolvation potentials (ND = −0.0098 kcal/mol/Å² and −0.0100 kcal/mol/Å²) (Figure 11). In the simulation with the default scaling factor (ND = −0.0090 kcal/mol/Å²), the low q-shoulder is shifted towards higher q-values and is increased in intensity. At q-values below 0.7 nm^{-1}, the structure factors obtained from mcMC simulations are lower than the experimental structure factors, indicating some discrepancy in the long-range order. This discrepancy could be due to the increased long-ranged repulsion caused by the additional charge, and/or to

the lack of convergence of the values of the radial distribution functions at large distances that are necessary to compute accurate values of $S(q)$ at very low q.[4] Nonetheless, the simultaneous overall good agreement of the mcMC simulation results (for a particular choice of the nonpolar desolvation scaling factor, ND = -0.0098 kcal/mol/Å^2) with experimentally determined osmotic second virial coefficients and structure factors shows that the mcMC simulations can provide a realistic description of protein-protein interactions. Moreover, the conformational sampling provided by mcMC alleviates the unwanted dependence of the simulated protein-protein interactions on the choice of input structure, which is an issue with conventional rigid-body simulations.

High concentrations and the resulting prevalence of protein-protein interactions also affect the population of HEWL conformations in the mcMC simulations. Conformations that are able to form dimers or oligomers with low intermolecular potential energies are stabilized, while conformations that interact less favorably with other proteins are destabilized. Figure 12 shows the distribution of conformations sampled by the MD simulation represented by the 50 individual conformations in the library (gray bars) together with interaction-induced changes observed in the mcMC simulations at 169 mg/mL for the three values of the nonpolar desolvation potential considered here (colored bars). Both, stabilization and destabilization effects are monotonic and approximately proportional to the scaling factor of the nonpolar desolvation potential. This result shows how mcMC simulations allow the system to adapt the distribution of protein conformations from dilute-limit conditions, in which the structural library was generated, to the

high concentration regime, in which inter-protein interactions become relevant. This adaptation is limited to a change in the population of discrete conformations represented in the employed library of structures. Conformations that are unfavorable under dilute conditions and only become stable due to interactions with other proteins will have low statistical weights in a library generated by a MD simulation of a single solvated protein and, hence, require many trial moves to be sampled, even if they are energetically favored in high concentration conditions. Alternative procedures for generating the library of protein conformations, such as a more computationally demanding MD simulation of multiple protein molecules at high concentration, might be considered in such a case.

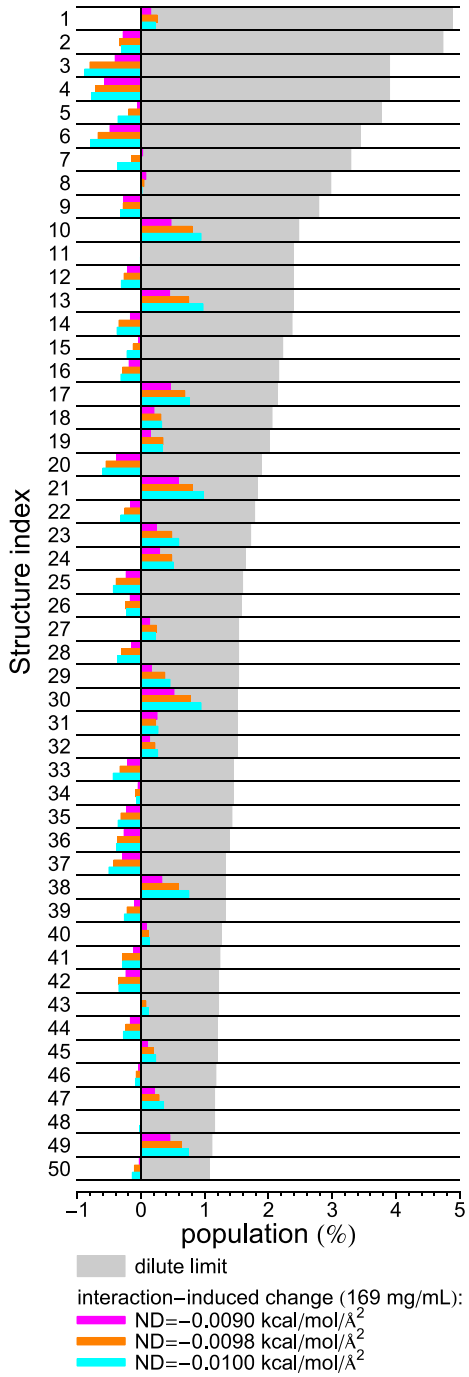


Figure 12. Distribution of protein conformations sampled by the MD simulation (gray bars, proportional to the statistical weights shown in Figure 8B) together with changes induced by inter-protein interactions in mcMC simulations at high protein concentration

(169 mg/mL) for varying strength of the nonpolar desolvation potential term (colored bars).

CHAPTER 3. SUMMARY

We have shown that MC simulations of many-protein systems can be performed using protein-protein interaction models developed for BD simulations. We used scMC simulations to analyze the effects of fixed side chain conformations in rigid-body simulations of systems containing many interacting proteins. Differences in the exposure of charged basic side chains, in particular between solution NMR structures and structures obtained from crystallography, can significantly modify preferential protein-protein interaction sites and the overall attraction of the proteins. Within the framework of MC simulations of protein solutions, we introduced a simple approach, mcMC, to account for molecular flexibility by allowing molecules to switch between multiple conformations within a discrete library of conformations, e.g., as obtained herein from a MD simulation of a single, fully-flexible protein in aqueous solution. Statistical weights used in the generation of mcMC trial moves for conformational changes allow us to describe distinct thermodynamic stabilities of individual conformations in the infinite dilution limit. Our approach removes potential artifacts observed in simulations of rigid protein structures, such as highly specific binding motifs involving fixed conformations of long, charged side chains, whose flexibility needs to be accounted for. In particular, after a minor reparameterization of empirical scaling parameters in the protein-protein interaction potential, we demonstrated improved agreement of simulated osmotic second virial coefficients with light scattering experiments at low protein concentrations and various salt concentrations. In addition, we could demonstrate

improved agreement with experimental structure factors obtained at high protein concentration.

The mcMC approach can be employed for simulations of aggregating protein systems and crowded biomolecular solutions. Introducing flexibility for the simulated proteins, even if only within a limited set of discrete, Boltzmann-weighted conformations, will improve predictions of specific binding modes as well as overall aggregation propensities. Furthermore, the use of MC simulations allows implementation of enhanced sampling procedures for strongly aggregating systems (such as aggregation volume biased Monte Carlo[10, 11, 13, 22]), which are otherwise challenging to sample in conventional BD or MC simulations.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health (grant R01 EY021514 to R.W.M.), the National Science Foundation (grants DMR-1410415 to R.W.M. and D.J.T and DMS-1361425 to C.T.B. and R.W.M.), and the Cluster of Excellence RESOLV (EXC 1069) funded by the Deutsche Forschungsgemeinschaft (M.H.). M.H. was supported by a fellowship from the German Academy of Sciences Leopoldina. We are grateful to Professor Stephen White in the Department of Physiology and Biophysics at UC Irvine for allowing us to use his instrument for the light scattering measurements, and to Dr. Adrian Elcock and Dr. Rebecca Wade for providing their Brownian dynamics codes and helpful advice.

REFERENCES

1. Ellis, R. J. Macromolecular Crowding: Obvious but Underappreciated. *Trends Biochem. Sci.* **2001,** *26*, 597-604.
2. Ellis, R. J.; Minton, A. P. Protein Aggregation in Crowded Environments. *Biol. Chem.* **2006,** *387*, 485-497.
3. Feig, M.; Sugita, Y. Variable Interactions between Protein Crowders and Biomolecular Solutes Are Important in Understanding Cellular Crowding. *J. Phys. Chem. B* **2012,** *116*, 599-605.
4. McGuffee, S. R.; Elcock, A. H. Atomically Detailed Simulations of Concentrated Protein Solutions: The Effects of Salt, pH, Point Mutations, and Protein Concentration in Simulations of 1000-Molecule Systems. *J. Am. Chem. Soc.* **2006,** *128*, 12098-12110.
5. McGuffee, S. R.; Elcock, A. H. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *Plos Comp. Biol.* **2010,** *6*, e1000694.
6. Mereghetti, P.; Gabdoulline, R. R.; Wade, R. C. Brownian Dynamics Simulation of Protein Solutions Structural and Dynamical Properties. *Biophys. J.* **2010,** *99*, 3782-3791.
7. Dlugosz, M.; Trylska, J. Diffusion in Crowded Biological Environments: Applications of Brownian Dynamics. *BMC Biophys.* **2011,** *4*, 3.
8. Ermak, D. L.; Mccammon, J. A. Brownian Dynamics with Hydrodynamic Interactions. *J. Chem. Phys.* **1978,** *69*, 1352-1360.
9. Marcos, E.; Mestres, P.; Crehuet, R. Crowding Induces Differences in the Diffusion of Thermophilic and Mesophilic Proteins: A New Look at Neutron Scattering Results. *Biophys. J.* **2011,** *101*, 2782-2789.
10. Chen, B.; Siepmann, J. I. A Novel Monte Carlo Algorithm for Simulating Strongly AssociatingFfluids: Applications to Water, Hydrogen Fluoride, and Acetic Acid. *J. Phys. Chem. B* **2000,** *104*, 8725-8734.
11. Chen, B.; Siepmann, J. I. Improving the Efficiency of the Aggregation-Volume-Bias Monte Carlo Algorithm. *J. Phys. Chem. B* **2001,** *105*, 11275-11282.
12. Lomakin, A.; Asherie, N.; Benedek, G. B. Monte-Carlo Study of Phase Separation in Aqueous Protein Solutions. *J. Chem. Phys.* **1996,** *104*, 1646-1656.
13. Chen, B.; Nellas, R. B.; Keasler, S. J. Fractal Aggregates in Protein Crystal Nucleation. *J. Phys. Chem. B* **2008,** *112*, 4725-4730.
14. Staneva, I.; Frenkel, D. The Role of Non-Specific Interactions in a Patchy Model of Protein Crystallization. *J. Chem. Phys.* **2015,** *143*, 194511.
15. Liu, H.; Kumar, S. K. Vapor-Liquid Coexistence of Patchy Models: Relevance to Protein Phase Behavior. *J. Chem. Phys.* **2007,** *127*, 084902.
16. Fusco, D.; Charbonneau, P. Crystallization of Asymmetric Patchy Models for Globular Proteins in Solution. *Phys. Rev. E* **2013,** *88*, 012721.
17. Lund, M.; Jönsson, B. A Mesoscopic Model for Protein-Protein Interactions in Solution. *Biophys. J.* **2003,** *85*, 2940-2947.
18. Lund, M. Anisotropic Protein-Protein Interactions Due to Ion Binding. *Coll. Surf. B: Biointerfaces* **2015,** *137*, 17-21.

19. Li, W.; Persson, B. A.; Morin, M.; Behrens, M. A.; Lund, M.; Oskolkova, M. Z. Charge-Induced Patchy Attractions between Proteins. *J. Phys. Chem. B 119*, 503-508.
20. Kurut, A.; Persson, B. A.; Åkesson, T.; Forsman, J.; Lund, M. Anisotropic Interactions in Protein Mixtures: Self Assembly and Phase Behavior in Aqueous Solution. *J. Phys. Chem. Lett.* **2012,** *3*, 731-734.
21. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953,** *21*, 1087-1092.
22. Chen, B.; Siepmann, J. I.; Oh, K. J.; Klein, M. L. Aggregation-Volume-Bias Monte Carlo Simulations of Vapor-Liquid Nucleation Barriers for Lennard-Jonesium. *J. Chem. Phys.* **2001,** *115*, 10903-10913.
23. Errington, J. R.; Panagiotopoulos, A. Z. New intermolecular potential models for benzene and cyclohexane. *J. Chem. Phys.* **1999,** *111* (21), 9731-9738.
24. Macedonia, M. D.; Maginn, E. J. A biased grand canonical Monte Carlo method for simulating adsorption using all-atom and branched united atom models. *Mol. Phys.* **1999,** *96* (9), 1375-1390.
25. Sepehri, A.; Loeffler, T. D.; Chen, B. Improving the efficiency of configurational-bias Monte Carlo: A density-guided method for generating bending angle trials for linear and branched molecules. *J. Chem. Phys.* **2014,** *141* (7).
26. Shah, J. K.; Maginn, E. J. A general and efficient Monte Carlo method for sampling intramolecular degrees of freedom of branched and cyclic molecules. *J. Chem. Phys.* **2011,** *135* (13).
27. Gabdoulline, R. R.; Wade, R. C. On the Contributions of Diffusion and Thermal Activation to Electron Transfer between Phormidium laminosum Plastocyanin and Cytochrome f: Brownian Dynamics Simulations with Explicit Modeling of Nonpolar Desolvation Interactions and Electron Transfer Events. *J. Am. Chem. Soc.* **2009,** *131*, 9230-9238.
28. Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *J. Am. Chem. Soc.* **1988,** *110*, 1657-1666.
29. Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R., et al. Electrostatics and Diffusion of Molecules in Solution: Simulations with the University of Houston Brownian Dynamics Program. *Comp. Phys. Commun.* **1995,** *91*, 57-95.
30. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. USA* **2001,** *98*, 10037-10041.
31. Gabdoulline, R. R.; Wade, R. C. Effective Charges for Macromolecules in Solvent. *J. Phys. Chem.* **1996,** *100*, 3868-3878.
32. Niimura, N.; Minezaki, Y.; Nonaka, T.; Castagna, J. C.; Cipriani, F.; Hoghoj, P.; Lehmann, M. S.; Wilkinson, C. Neutron Laue Diffractometry with an Imaging Plate Provides an Effective Data Collection Regime for Neutron Protein Crystallography. *Nat. Struct. Biol.* **1997,** *4*, 909-914.
33. Bon, C.; Lehmann, M. S.; Wilkinson, C. Quasi-Laue Neutron-Diffraction Study of the Water Arrangement in Crystals of Triclinic Hen Egg-White lLysozyme. *Acta Cryst. D* **1999,** *55*, 978-987.

34. Schwalbe, H.; Grimshaw, S. B.; Spencer, A.; Buck, M.; Boyd, J.; Dobson, C. M.; Redfield, C.; Smith, L. J. A Refined Solution Structure of Hen Lysozyme Determined Using Residual Dipolar Coupling Data. *Protein Sci.* **2001,** *10*, 677-688.
35. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem.-Intl. Ed.* **1999,** *38*, 236-240.
36. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory and Comp.* **2008,** *4*, 435-447.
37. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **1983,** *79* (2), 926-935.
38. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995,** *103*, 8577-8593.
39. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a Linear Constraint Solver for Molecular Simulations. *J. Comp. Chem.* **1997,** *18*, 1463-1472.
40. Miyamoto, S.; Kollman, P. A. Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J. Comp. Chem.* **1992,** *13*, 952-962.
41. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984,** *81*, 3684-3690.
42. Nosé, S. A Molecular-Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984,** *52*, 255-268.
43. Parrinello, M.; Rahman, A. Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. *J. Appl. Phys.* **1981,** *52*, 7182-7190.
44. Hill, T. L., *An Introduction to Statistical Thermodynamics*. Dover Publications: New York, 1986; p xiv, 508 p.
45. Velev, O. D.; Kaler, E. W.; Lenhoff, A. M. Protein Interactions in Solution Characterized by Light and Neutron Scattering: Comparison of Lysozyme and Chymotrypsinogen. *Biophys. J.* **1998,** *75*, 2682-2697.
46. Zimm, B. H. The Scattering of Light and the Radial Distribution Function of High Polymer Solutions. *J. Chem. Phys.* **1948,** *16*, 1093-1099.
47. Wyatt, P. J. Light Scattering and the Absolute Characterization of Macromolecules. *Anal. Chem. Acta* **1993,** *272*, 1-40.
48. Krigbaum, W. R.; Kuegler, F. R. Molecular Conformation of Egg-White Lysozyme and Bovine alpha-Lactalbumin in Solution. *Biochemistry* **1970,** *9*, 1216--1223.
49. *R: a Language and Environment for Statistical Computing*; R Development Core Team: Vienna, Austria, 2008.
50. Efron, B.; Tibshirani, R., *An Introduction to the Bootstrap*. Chapman and Hall: London, 1993.
51. Piazza, R.; Pierno, M. Protein Interactions Near Crystallization: a Microscopic Approach to the Hofmeister Series. *J. Phys.-Cond. Matter* **2000,** *12*, A443-A449.
52. Muschol, M.; Rosenberger, F. Interactions in Undersaturated and Supersaturated Lysozyme Solutions: Static and Dynamic Light Scattering Results. *J. Chem. Phys.* **1995,** *103*, 10424-10432.

53. Guo, B.; Kao, S.; McDonald, H.; Asanov, A.; Combs, L. L.; Wilson, W. W. Correlation of Second Virial coefficients and Solubilities Useful in Protein Crystal Growth. *J. Cryst. Growth* **1999,** *196*, 424-433.
54. Rosenbaum, D. F.; Kulkarni, A.; Ramakrishnan, S.; Zukoski, C. F. Protein Interactions and Phase Behavior: Sensitivity to the Form of the Pair Potential. *J. Chem. Phys.* **1999,** *111*, 9882-9890.
55. Pellicane, G.; Costa, D.; Caccamo, C. Microscopic Determination of the Phase Diagrams of Lysozyme and gamma-Crystallin Solutions. *J. Phys. Chem. B* **2004,** *108*, 7538-7541.
56. Kuehner, D. E.; Engmann, J.; Fergg, F.; Wernick, M.; Blanch, H. W.; Prausnitz, J. M. Lysozyme Net Charge and Ion Binding in Concentrated Aqueous Electrolyte Solutions. *J. Phys. Chem. B* **1999,** *103*, 1368-1374.
57. Stradner, A.; Cardinaux, F.; Schurtenberger, P. A Small-Angle Scattering Study on Equilibrium cCusters in Lysozyme Solutions. *J. Phys. Chem. B* **2006,** *110*, 21222-21231.
58. Liu, Y.; Fratini, E.; Baglioni, P.; Chen, W. R.; Chen, S. H. Effective Long-Range Attraction Between Protein Molecules in Solutions Studied by Small Angle Neutron Scattering. *Phys. Rev. Lett.* **2005,** *95*, 118402.

SUPPLEMENTAL INFORMATION

Table S1. Single-conformation Monte Carlo (scMC) simulations of hen egg white lysozyme solutions

Protein Structure (PDB ID)	Net Charge (*e*)	Number of Proteins	Concentration (mg/ml)	Ionic Strength (mM)	MC cycles[a] (x 10^3)
1E8L	+9	200	10	50	1500
1E8L	+9	200	10	100	1500
1E8L	+9	200	10	150	1500
1E8L	+9	200	10	200	1500
1E8L	+9	200	10	250	1500
1E8L	+9	200	10	300	1500
1E8L	+9	200	10	500	1500
1E8L	+9	475	169	100	385
1IO5	+9	200	10	50	1500
1IO5	+9	200	10	100	1500
1IO5	+9	200	10	150	1500
1IO5	+9	200	10	200	1500
1IO5	+9	200	10	250	1500
1IO5	+9	200	10	300	1500
1IO5	+9	200	10	500	1500
1LZN	+11	200	10	50	1500
1LZN	+11	200	10	100	1500
1LZN	+11	200	10	150	1500
1LZN	+11	200	10	200	1500
1LZN	+11	200	10	250	1500
1LZN	+11	200	10	300	1500
1LZN	+11	200	10	500	1500

[a]One MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Table S2. Multi-conformation Monte Carlo (mcMC) simulations of hen egg white lysozyme solutions[a,b]

Number of Proteins	Concentration (mg/ml)	Ionic Strength (mM)	MC cycles[c] (x 10^3)
200	10	50	1500
200	10	100	1500
200	10	150	1500
200	10	200	1500
200	10	250	1500
200	10	300	1500
200	10	500	1500
475	169	50	300

[a]The structure library used in the mcMC simulations was generated from an all-atom MD simulation with the 1E8L structure (net charge +9e) as the initial configuration.
[b]Each simulation indicated in this table was performed three times, each with a different value of the nonpolar desolvation parameter.
[b]One MC cycle consists of N_{prot} trial moves, where N_{prot} is the number of protein molecules in the simulation.

Dependence of protein-protein radial distribution functions on the size of the interaction potential grids. All of the terms in the protein-protein interaction potential are mapped onto cubic grids for computational efficiency. We investigated the optimal grid size for each potential term in turn by varying the grid size between 60^3 and 200^3 Å^3, while using a 200^3 Å^3 grid for all the other potential terms. Figure S1 shows the corresponding radial distribution functions for each test set. Figure S1 shows radial distribution functions obtained from scMC simulations using the 1E8L NMR solution structure at 10 mg/mL concentration and 100 mM ionic strength. The results suggest that convergence is achieved for the electrostatic potential at a minimum grid size of 100^3 Å^3 (at this ionic strength) and at 80^3 Å^3 for all other potential terms. The conservative use of 200^3 Å^3 potential grids in the remainder of this study therefore ensures minimal influence of this effective interaction potential cutoff on the reported results.

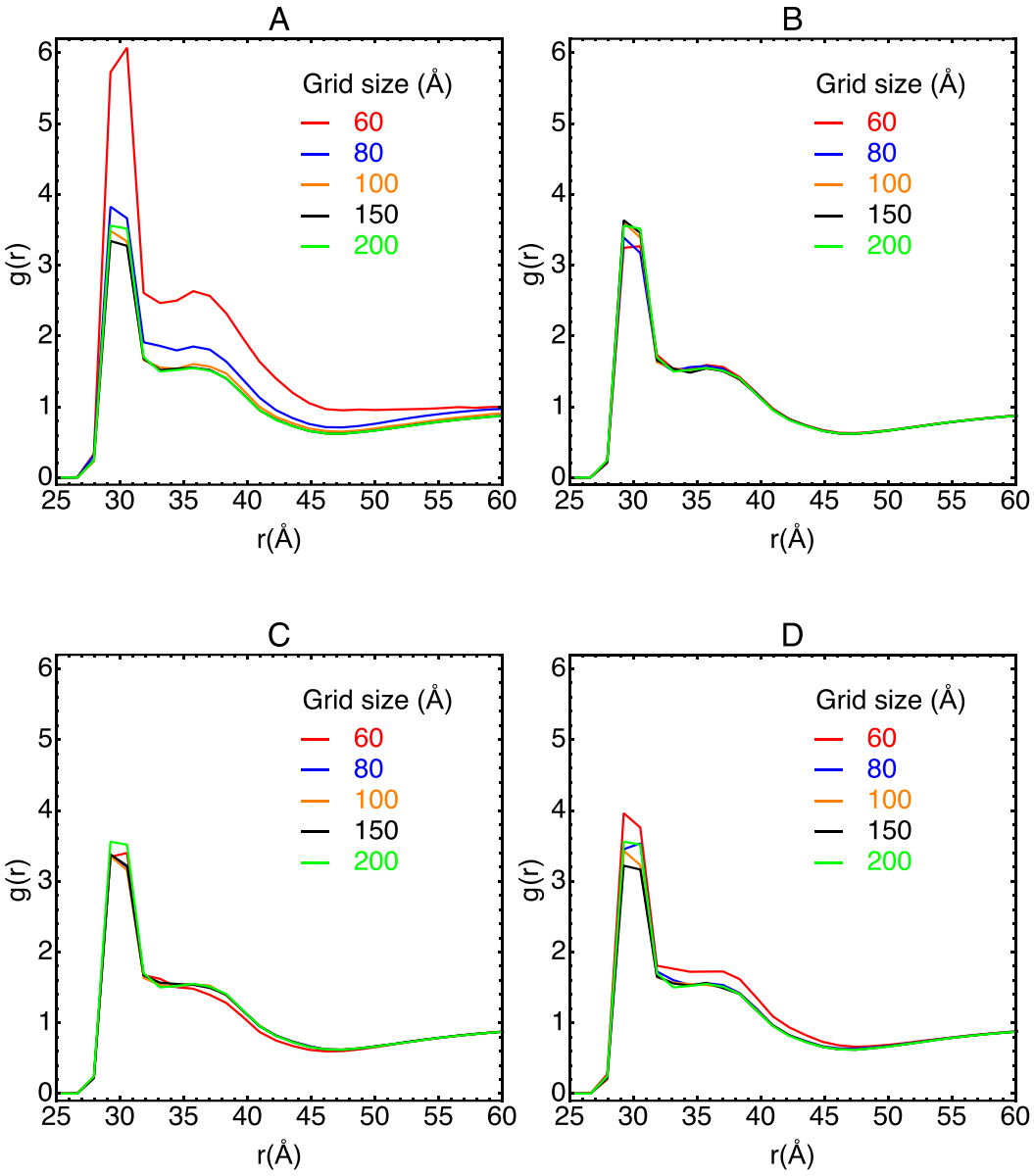
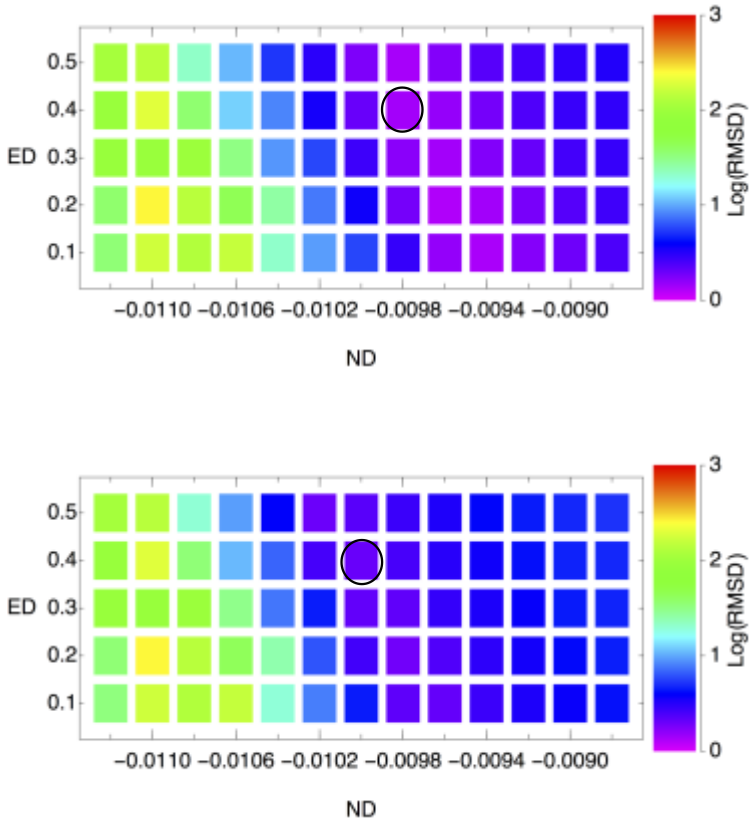


Figure. S1. Protein-protein radial distribution functions from HEWL scMC simulations performed with the 1E8L structure at a concentration of 10 mg/mL and an ionic strength of 100 mM. Each panel shows the effect of varying the grid size of a specific interaction potential term while using a grid size of at 200^3 $Å^3$ for all others: (A) Electrostatic potential term. (B) Electrostatic desolvation potential. (C) Nonpolar desolvation potential. (D) Soft-core repulsion potential.

Optimization of the parameters in the desolvation potentials. In the SDAMM potential,[1] the strength of the electrostatic and nonpolar desolvation potential terms are specified by empirical scaling paramters (ED and ND, respectively).[1, 2] As the ionic strength is increased, the Coulomb interactions are screened and the short-

ranged desolvation potentials dominate the protein-protein interactions. In order to optimize the values of ED and ND, we performed a systematic search of the ED/ND parameter space by computing B_2 values in HEWL mcMC simulations at a concentration of 10 mg/ml and ionic strengths of 100 and 300 mM. We calculated the root-mean squared differences (RMSD) between the simulation values and our two sets of experimental B_2 estimates (pH 4.7 and 6.9) in the same range of ionic strength. The ND parameter was varied from –0.0114 kcal/mol/Å^2 to –0.0088 kcal/mol/Å^2 and the ED parameter from 0.1 to 0.5. The logarithm of the RMSD is shown as a function of the ND and ED parameters in Figure S2. It is evident that the attractive and repulsive contributions to the overall protein-protein interaction potential due to the two desolvation potentials can compensate each other to some degree, hence the corresponding scaling parameters are not entirely independent.

Because the resulting set of optimal (ED,ND) pairs includes the scaling value of the electrostatic desolvation term employed in previous HEWL BD simulations[1] (ED = 0.36), we opted for leaving ED at this default value and explored the effect of changing ND in more detail. In addition to the ND value used in HEWL BD simulations[1] (ND = -0.090 kcal/mol/Å^2), we report in the main text mcMC simulation results for the two ND values corresponding to the minimum B_2 RMSDs for ED = 0.4 (ND = –0.098 kcal/mol/Å^2 when comparing to B_2 experimental estimates at pH 4.7; ND = –0.0100 kcal/mol/Å^2 RMSD when comparing to B_2



experimental estimates at pH 6.9).

Figure. S2. Logarithm of the root-mean squared differences (RMSD) between B_2 estimates from mcMC HEWL simulations (using the indicated desolvation potential parameters) and experimental estimates at pH 4.7 (top) and pH 6.9 (bottom) and ionic strengths of 100 mM and 300 mM. The optimal ND values (at ED = 0.4) are encircled.

REFERENCES FOR THE SUPPLEMENT

1. Mereghetti, P.; Gabdoulline, R. R.; Wade, R. C. Brownian Dynamics Simulation of Protein Solutions Structural and Dynamical Properties. *Biophys. J.* **2010,** *99*, 3782-3791.
2. Gabdoulline, R. R.; Wade, R. C. On the Contributions of Diffusion and Thermal Activation to Electron Transfer between Phormidium laminosum Plastocyanin and Cytochrome f: Brownian Dynamics Simulations with Explicit Modeling of Nonpolar Desolvation Interactions and Electron Transfer Events. *J. Am. Chem. Soc.* **2009,** *131*, 9230-9238.