Supporting material for

Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data

Po-chia Chen[1] and Jochen S. Hub[1]

[1] Institute for Microbiology and Genetics, Georg-August-University Göttingen, Göttingen, Germany

* E-mail: jhub@gwdg.de

Supporting Material Methods

Construction of solvation layer using the envelope

The solvation envelope is defined with respect to a reference structure in a fixed orientation ω, whose center of mass is located at the origin. Thus, the protocol involves the shifting of current system coordinates around the solute in orientation ω' (i), calculation of the correct rotation (ii) and translation (iii) offsets to bring current system frame to the reference frame at the origin (v), but packing the solvent (iii/iv) such that the entire envelope is filled correctly. The protocol can be decomposed into the following steps:

(i) If the solute is broken over the periodic boundary, it is made whole by minimizing the distance of all solute atoms to a central pre-selected atom, or the center-of-geometry of a small set of atoms (termed pbc-atom(s), where pbc denotes 'periodic boundary conditions').

(ii) The pure rotation matrix $R_{\omega' \to \omega}$ is obtained by fitting the solute in its current orientation ω', to a reference structure in fixed orientation ω.

(iii) The envelope is rotated into the frame orientation ω' using rotation $R^{-1}_{\omega' \to \omega}$. To ensure that the envelope fits into the simulation box, the smallest enclosing sphere around the rotated envelope is constructed using the Miniball algorithm by Gärtner [1], yielding the center $\mathbf{r}_{\mathrm{es}}(\omega')$ of the enclosing sphere. The enclosing sphere with its envelope is translated by $\mathbf{r}_{\mathrm{box}} - \mathbf{r}_{\mathrm{es}}(\omega')$ into the compact unit cell (Wigner-Seitz cell) such that the center of the sphere is at the center $\mathbf{r}_{\mathrm{box}}$ of the cell. A check is made with the envelope vertices - if any lie outside the cell, this signifies that the envelope cannot be fully filled without periodic images of solvent atoms, and the algorithm stops with an error message suggesting a larger simulation box. This prevents possible periodicity artifacts from capturing both a solvent atom and its periodic image inside an envelope.

(iv) Having made the above safety check, both solute and solvent in the system frame ω' are translated into the compact unit cell, centering the solute center-of-mass at $\mathbf{r}_{\mathrm{box}} - \mathbf{r}_{\mathrm{es}}(\omega')$. If the solute was successfully made whole in (i), this places all solute atoms into the rotated envelope. If not, some solute atoms will fall outside the cell and the algorithm will again stop with an error, and suggest better pbc-atoms selections. Solvent atoms outside of the compact unit cell are translated into the

compact cell by multiples of the box vectors. Step (iii) ensures that the envelope is completely filled by solvent after this step.

(v) Having passed the two checks above, all atoms are translated such that the center-of-mass of the solute is at the origin, and all atoms are rotated into the reference orientation ω using $R_{\omega' \to \omega}$.

(vi) Solvent atoms inside of the envelope are selected as the solvation layer. To avoid that this selection scales with $N_{\text{solvent}} \times N_{\text{faces}}$, where N_{solvent} denotes the number of solvent atoms and N_{faces} the number of faces of the envelope, we implemented a recursive algorithm that scales instead as $N_{\text{solvent}} \times (20 + 4M)$, where M is the number of recursions while constructing the icosphere (see main text).

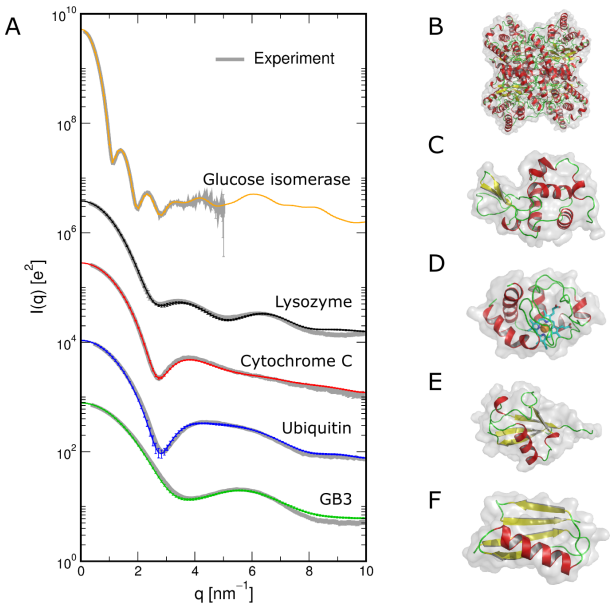

Figure S1. Comparison between calculated and experimental SAXS/WAXS patterns by a weighted fit. Here, the experimental curves were fitted to the calculated curves by minimizing $\chi^2_w = N_q^{-1} \sum_{i=1}^{N_q} \{[I_c(q_i) - f\, I_{\mathrm{exp}}(q_i) + c]/\sigma_{\mathrm{exp}}(q_i)\}^2$, where $\sigma_{\mathrm{exp}}(q_i)$ denote the experimental statistical errors. For clarity, the intensities for the five proteins were multiplied by (from top to bottom) constants of 10, 1, 0.1, 0.01, and 0.001, respectively.

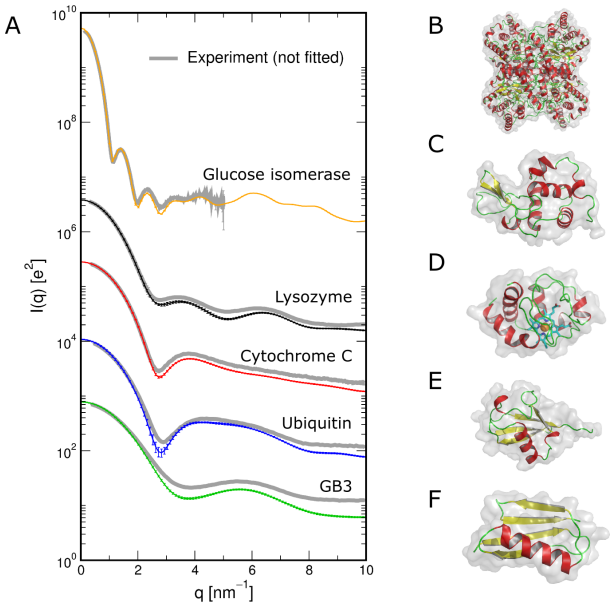

Figure S2. Comparison between calculated and experimental SAXS/WAXS patterns. Here, only the overall scale of the experimental spectra (grey curves) were fitted to the calculated spectra (colored curves) in the $q < 2\,\mathrm{nm}^{-1}$ range using a non-weighted fit. The slight discrepancy can be explained by uncertainties in the buffer subtraction and/or by dark currents, and it can be approximately absorbed by a single fitting parameter, leading to Figure 2 of the main text. For clarity, the intensities for the five proteins were multiplied by (from top to bottom) constants of 10, 1, 0.1, 0.01, and 0.001, respectively.

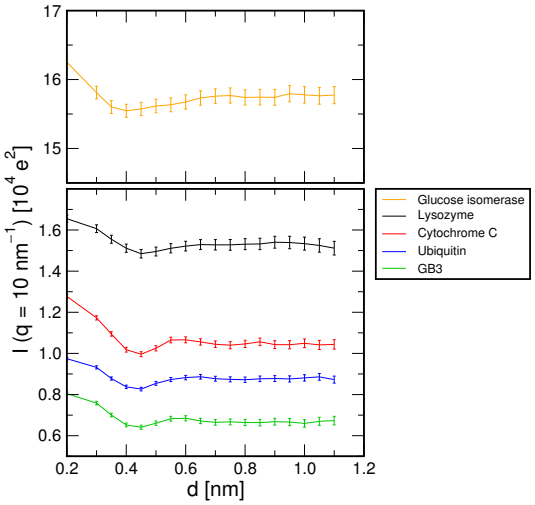

Figure S3. Effect of solvation layer on calculated curves at wide angles Net intensity at $q = 10\,\mathrm{nm}^{-1}$ versus the distance d between protein and envelope. In line with the findings for $I(0)$ and the radius of gyration (see main text), the intensity converges at $d > 0.7\,\mathrm{nm}$

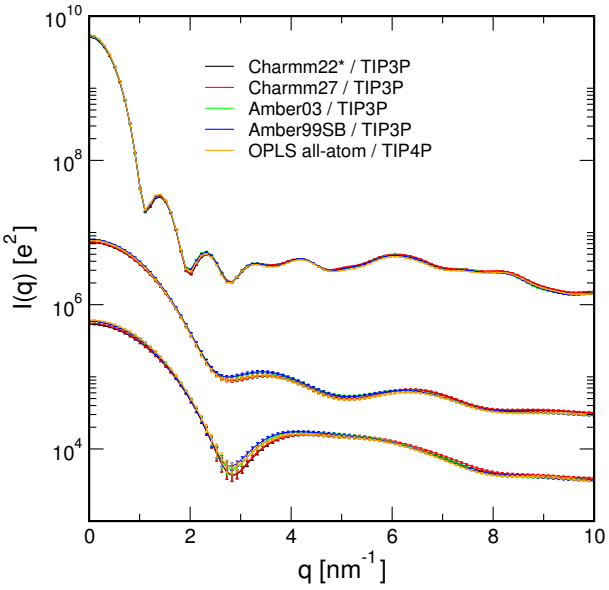

Figure S4. Spectra are independent of the protein force field in position-restrained simulations. WAXS spectra of (top to bottom) glucose isomerase, lysozyme, and ubiquitin computed from simulations using five different all-atom force fields, as indicated in the legend. Simulations were conducted for 10 ns with position restraints on the backbone atoms ($k = 1000\,\mathrm{kJ\,mol^{-1}nm^{-2}}$). The spectra agree within statistical errors. 1000 frames were used for the average.

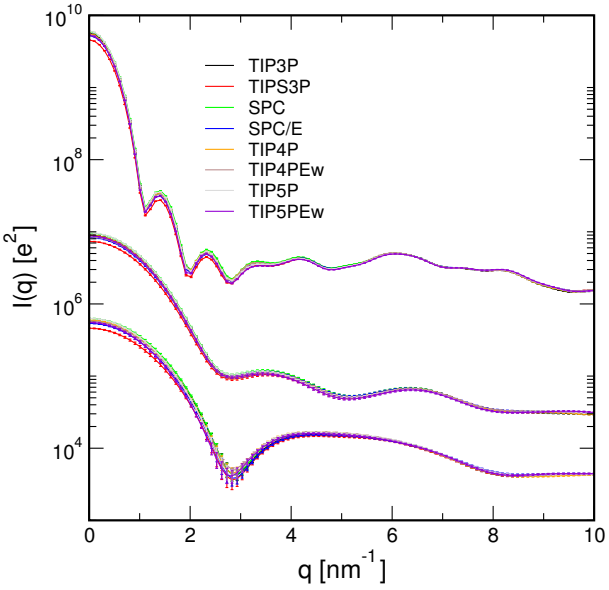

Figure S5. SAXS and near-WAXS patterns of glucose isomerase, lysozyme, and ubiquitin calculated using eight different water models and without density correction for the bulk solvent. Variations due to the density of the water model by up to 30% appear. These variations are reduced to ∼5% upon the application of a density correction.

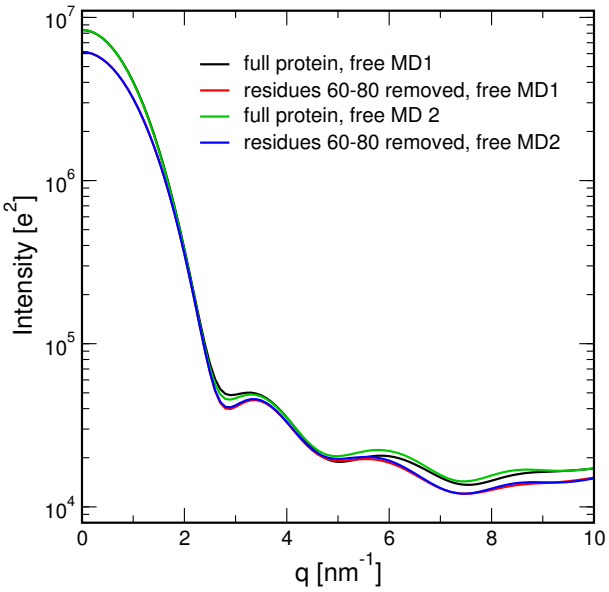

Figure S6. Role of the loop on intensity differences between free lysozyme simulations MD 1 and MD 2. To validate that the improved fitting of the free MD 1 as compared to the free MD2 is due to the loop between residues 60 and 80, we computed SWAXS curves from (i) the complete protein and (ii) after deleting the loop at residues 60-80. For that purpose, 22 snapshots were taken from each of the simulations MD 1 and MD 2. SWAXS intensities were computed using CRYSOL and averaged over those 22 snapshots. The intensities calculated after removing the loop are nearly indistinguishable between MD 1 and MD 2 (red and blue curve) suggesting that core structure of lysozyme was similar during MD 1 and MD 2. In contrast, intensities calculated from the full protein differ (black and green), confirming a small yet significant influence of the loop.

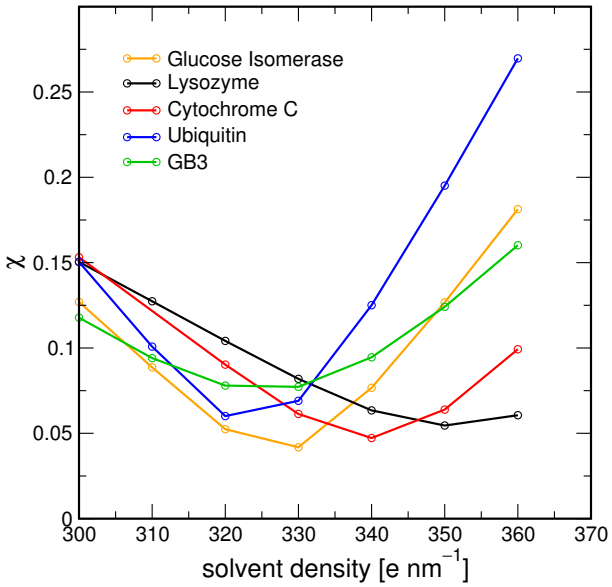

Figure S7. Agreement between calculated and experimental WAXS curves as a function of preselected solvent density. To exclude the possibility that fitting a constant parameter c (eq. 13) would absorb alterations in the solvent density, we computed the WAXS curves over a range of specified solvent densities (see paragraph on Solvent density correction). For most proteins, χ is highly sensitive to these variations, suggesting the fitting procedure does not absorb unphysical buffer densities.

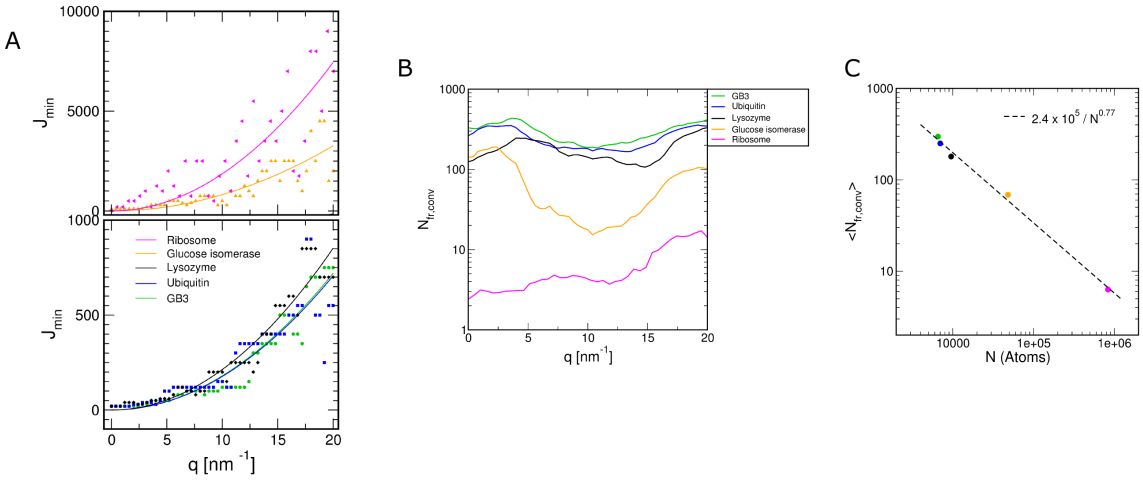

Figure S8. On the required number J of q-vectors for the orientational average and required number of simulation frames.
(A) Required minimum number $J_{\min}$ to achieve an accuracy of 2% for the orientational average. $J_{\min}$ was defined as the smallest value that fulfills $|I(q;J) - I_{\text{conv}}(q)| < \Delta I(q)$ for all $J > J_{\min}$. Here, $I_{\text{conv}}(q)$ is the nearly fully converged intensity that was computed using $J = 20000$. For the acceptable uncertainty $\Delta I(q)$ we chose $0.02\,I(q)$ if $q < 5\,\text{nm}^{-1}$. For $q > 5\,\text{nm}^{-1}$, where the SWAXS curves do not decay as rapidly as for small angles, we chose $\Delta I(q) = 0.02 I^{(\text{av})}_{\text{waxs}}$, where $I^{(\text{av})}_{\text{waxs}}$ denotes the average of $I(q)$ between 5 and $10\,\text{nm}^{-1}$. The lines are fitted parabolas $J_{\min} = \alpha(Dq)^2$, where D is the maximum diameter of the respective envelope. For the five biomolecules shown here (from top to bottom) the fits yield $\alpha = 0.04$, 0.06, 0.04, 0.05, and 0.06, respectively. Hence, the estimate for α is consistent. To ensure an accurate average also at very small angles, however, we suggest to add a constant offset to J, such as $J(q) = \alpha(Dq)^2 + 100$.
(B) Approximate required number of simulation frames $N_{\text{fr,conv}}$ to converge the SWAXS curves from position restrained simulations to 2% accuracy. $N_{\text{fr,conv}}$ was chosen such that, for all $N_{\text{fr}} > N_{\text{fr,conv}}$, $|I(q;N_{\text{fr}})/I(q;N^{(\max)}_{\text{fr}}) - 1| < 0.02$ *or* $|I(q;N_{\text{fr}}) - I(q;N^{(\max)}_{\text{fr}})| < 0.02 I^{(\text{av})}_{\text{waxs}}$. Here, $I(q;N_{\text{fr}})$ denotes the intensity calculated from the first N_{fr} simulation frames. $I^{(\text{av})}_{\text{waxs}}$ is the average intensity between 5 and $10\,\text{nm}^{-1}$ calculated from $N^{(\max)}_{\text{conv}}$ simulation frames (100 for the ribosome, and 1000 for all other solutes). Because $N_{\text{fr,conv}}$ is highly noisy if extracted from a single WAXS calculation, $N_{\text{fr,conv}}$ was averaged over the calculations of 8 different water models, and subsequently slightly smoothed along q. Figure B demonstrates that the required number of frames drastically decrease with the size of the solute.
(C) $N_{\text{fr,conv}}$ averaged over the entire q-range of Figure B versus the number N of atoms in the envelope. The fitted line sugget that the required number of frames to achieve a converged WAXS curve decreases approximately as $N^{-0.77}$.

Table S1. Effect of atomic fluctuations and conformational sampling on WAXS profiles: weighted fit

	frozen[a]	posres heavy at.[b]	posres backbone[c]	free MD 1	free MD 2	NMR ensemble
Glucose Isomerase	4.7	4.2	4.5	14.3		
Lysozyme	1.9	1.5	1.7	2.3	3.2	3.7[e]
Cytochrome-C	2.3	2.1	1.2	1.7		
Ubiquitin	6.7	4.7	4.8±0.6[d]	3.4	3.6	4.6[f] / 4.3[g]
GB3	2.7	3.2	2.7±0.2[d]	3.6	3.5	

Agreement between experimental and calculated SWAXS curves determined through a fit weighted by statistical experimental errors, $\chi^2_w = N_q^{-1} \sum_{i=1}^{N_q} \{[I_c(q_i) - f\, I_{\rm exp}(q_i) + c]/\sigma_{\rm exp}(q_i)\}^2$. [a]All protein coordinates frozen; [b]position restraints on all heavy atoms ($k = 2000\,{\rm kJ\,mol^{-1}nm^{-2}}$), [c]position restraints on backbone atoms only ($k = 2000\,{\rm kJ\,mol^{-1}nm^{-2}}$); [d]average and standard deviation from 10 independent simulations; [e,f,g]NMR ensembles 1E8L, 1XQQ, and 1D3Z, respectively.

Supporting References

1. Gärtner B (1999) Fast and robust smallest enclosing balls. In: Algorithms-ESA99, Springer. pp. 325–338.