Supplementary Information: The Effect of Hydration and Dynamics on the Mass Density of Single Proteins

Cameron C. W. McAllister [1], Lucas S. P. Rudden [2], Elizabeth H. C. Bromley [1]*, Matteo T. Degiacomi [1,3,4]*
[1] *Department of Physics, Durham University, Durham, UK*
[2] *Institute of Bioengineering, École polytechnique fédérale de Lausanne, Lausanne, Switzerland*
[3] *EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, UK*
[4] *School of Informatics, University of Edinburgh, Edinburgh, UK*
* *Corresponding authors: matteo.degiacomi@ed.ac.uk, e.h.c.bromley@durham.ac.uk*

Supplementary Methods

Protein dataset sequence alignment

Sequence alignment was used, as implemented in Biopython, to assess to similarity between proteins of the protein dataset, as well as to compare the sequences of BPTI and titin to those of this dataset. The `pairwise2.align.localxx` function was used, finding the optimal subsequence matching between each pair of sequences. The returned score, with one point awarded for each matching residue, is converted with respect to the length of the comparison sequence into a percentage to allow comparison between the alignment of different sequences (see Figure S2).

Data Analysis implementation details

We extract molecular dynamics simulation snapshots in the form of PDB files, and for each of them we instantiate an MDAnalysis Universe object [1–3]. To calculate the Radial Distribution Function (RDF) of water around each protein, we use 75 bins between 1.5 and 4.0 Å, with RDF values normalised against the volume of each bin. To calculate the mass density including all atoms with varying distances from the protein, we calculate the densities at 41 equally spaced distances between 1.5 and 4.0 Å. To accurately determine the position of the first minimum in the RDF (and hence the thickness of the first hydration shell) we perform a cubic interpolation of the region of interest and find its minimum (using the `interp1d` and `argrelextrema` methods within the Scipy library library[4], see example in Figure S25.

Definition of atomic radii

Protein volumes calculated are derived by either (a) including voxels for which the centres are closer to the centres of the nearest protein atom than the nearest water atom as part of the protein (i.e., ignoring the van der Waals radii) or (b) subtracting the van der Waals radii from the distances to each atom before calculating whether a voxel should be considered as part of the protein. Method (b) has two variants, with the van der Waals radii either being assumed on an elemental basis from standard data, or extracted from a force field file (here Amber ff14SB). We found that the absolute percentage difference between the volumes calculated with these two methods is only 0.33%, with a standard deviation of 0.28%, see Figures S6 and S7. Only 10 PDBs have a difference in calculated volume greater than 1%. A 1 Å step size was used for these calculations. Across 20 proteins with 3 randomly rotated repeats for each protein, including the effects of van der Waals radii lead to an 18% increase in computational time.

Leeway parameter

The leeway parameter is the extra distance that is added to the extremes of the protein position in each Cartesian dimension to define the boundaries of the grid surrounding the protein over which volume calculations take place. Very small leeway values (< 2 Å) mean that the grid will cut off water molecules surrounding the extreme positions of the protein and will encroach on the volume surrounding the points representing protein atom positions at these locations. This results in a reduced value for the protein volume (and therefore an increased value for protein density). Hence, for accurate volume calculations the leeway should be at least large enough to accommodate the hydration shell of water molecules surrounding a protein. In practice, given the regions where a protein is close to the extremes along any one dimension are small, leeway effects tend to be negligible for values greater than 1 Å. We note that the chosen leeway value also effectively shifts the position that the protein occupies relative to the grid, which in turn changes the calculated protein volume. This effect becomes negligible for sufficiently small step sizes (below 1 Å).

Step size parameter

To establish a ground truth which future calculations of protein volume could be compared against, we performed two tests using a single simulation frame from a 500 ns simulation of BPTI (PDB: 5PTI, see Figure 5). Firstly, we per-

formed 1,000 random rotations with a 1.4 Å step size each averaged around 10 different axes of rotation to obtain a value of 8279.4 ± 0.3 Å^3. Secondly, we calculated the volume from the same simulation frame using an extremely fine grid size of 0.05 Å, for which we obtained a value of 8277.6 Å^3. Furthermore, we conducted further tests using 5 randomly chosen PDBs from the protein dataset (see Figure S3) to test the effect of varying the step size and leeway parameters on a broader range of protein structures. The ground truths for these protein structures are taken as the results for a 0.3 Å step size averaged over 10 rotations, and the absolute errors are calculated relative to this value. A step size of 0.5 Å with no rotations consistently reproduces the ground truth to within 0.2%, while having a significantly lower calculation duration (see Figure S5).

Effect of random rotations

The `transformations.rotate.rotateby` function of the `MDAnalysis.transformations` submodule of MDAnalysis[5] can be used to rotate a protein. The pseudo-random number generating functions from the NumPy[6] library are used to generate a random axis and degree of rotation. By rotating the protein, the position of the default 3-dimensional grid is altered relative to the protein position. Thus, averaging over multiple rotations improves the accuracy of the protein volume calculated via our grid-based method.

Figures 4 and 5 show that increasing the number of rotations generally improves the accuracy of the protein volume calculated via our method, while also increasing calculation time (linearly with an increasing number of rotations). However, high accuracy can also be achieved without rotations by using a small step size. Depending on the accuracy required this can be favourable in terms of computation runtime. Based on our benchmarks, we chose a step size of 0.5 Å and no rotations.

Identification of surface and interior atoms

The Solvent Accessible Surface Area (SASA) algorithm implemented in the Biobox Python package [7] returns a list of indices for surface atoms, defined as atoms for which a certain threshold fraction of mesh points surrounding each atom are found to be accessible to a probe of a certain radius during the Shrake-Rupley rolling ball method calculation of the Solvent Accessible Surface Area (SASA). In this work, we set a threshold of 5% and a probe radius of 1.4 Å. Figure S8 shows a visualisation of surface atoms identified in the extracellular domain of tissue factor (PDB: 1AHW). Conversely, atoms which are not exposed to solvent are classified as internal atoms.

Amino acids classification

The general classifications of amino acids used in this work are shown in table III. Proline, Tyrosine, and Cysteine, which are alternatively defined as hydrophobic or otherwise depending on the exact experiment[8] are excluded from the definition of hydrophobic residues.

Markov State Modelling

Markov State Modelling (MSM) was performed on a down-sampled 1 ms BPTI simulation with 103125 frames and a timestep of 10 ns. Model determination and validation are presented in Figure 23. We used a lag of 20 steps (200 ns) and 3 dimensions for the Time-lagged Independent Component Analysis (TICA) and 300 clusters for the κ-means clustering, which were found by iterating after inspection of the implied timescales and the associated Chapman-Kolmogorov tests. A final Chapman-Kolmogorov test was used to validate our choice of parameters. From our final MSM, we determined the transition matrix of our final MSM and mean first passage times.

Random forest regression

Random forest regressors were grid optimized using the `GridSearchCV function` of the `sklearn.model_selection` submodule, with test parameters of `max_depth = [2, 3, 5, 7, 10]`, `min_samples_leaf = [1, 2, 3, 4, 5]` and `n_estimators = [10, 15, 20, 30, 40, 50]`. For the structure-based random forest values of 10, 4, and 50 were found for the max tree depth, the minimum number of samples per leaf, and the number of decision trees, respectively. For the sequence-based random forest, grid optimization gave values of 5, 1, and 20.

We used permutation feature importance to quantify the importances of each feature to our final struct-RFR and seq-RFR models, with 10 permutation repeats for each feature conducted for each model using the `permutation_importance` function of the `sklearn.inspection` submodule.

To evaluate the performances of different random forest training approaches we performed 30 repeats for each set of random forest training data, each using 90% of the proteins in the protein dataset, to compare statistics including the Adjusted-R^2 and the Pearson's Correlation Coefficient for training and testing of these random forest regressors (see Table IV and Figure S17). While using structural characteristics and sequences of the full 5460 structures (21 simulation frames from each protein of the protein dataset) gives the best results when training on 90% of the proteins (Adjusted R^2: 0.945 ± 0.001, PCC: 0.9723 ± 0.0004), it is poorer when

tested on the remaining 10% of the proteins (Adjusted R^2: 0.50 $\pm$ 0.08, PCC: 0.75 $\pm$ 0.04) than a model trained using the mean structure-derived characteristics and sequence of each of the proteins (Adjusted R^2: 0.61 $\pm$ 0.06, PCC: 0.83 $\pm$ 0.03). This implies that training using multiple similar structures of each protein leads to over-fitting of the random forest regressor, reducing its ability to predict the densities of proteins outside the training set, which has implications for future training strategies. Both models trained using structural characteristics outperform the model trained using only amino acid sequences in both training and testing, though the testing Pearson's Correlation Coefficient of the sequence random forest is still strong, at 0.72 $\pm$ 0.04.

Sphericity calculation

For sphericity calculations a probe radius of 3.0 Å, found by Kim et al. to be the best radius for determining sphericity [9], was used in the calculation of the protein volume using ProteinVolume and the SASA using Biobox, as previous. Protein sphericity was then calculated as:

$$S = \frac{\pi^{1/3}(6V)^{2/3}}{A}, \tag{1}$$

where S is the sphericity, V is the protein volume, and A is the SASA. A sphericity of 1 indicates a perfect sphere, 0 an infinite plane.

Analysis of water structure and order

To analyse the structure of water we utilise three order parameters: the orientational tetrahedral order parameter[10] (q), the translational tetrahedral order parameter[11] (S_k), and the local structure index (LSI), see Figure S26.

The orientational tetrahedral order parameter, q, was first proposed by Chau and Hardwick in 1998[10], with Errington and Debenedetti[12] later adding a scaling factor ensuring that q varies between 0 for an ideal gas to 1 for a regular tetrahedron. It is calculated with respect to the four nearest water oxygen atoms to the central water oxygen atom under consideration and is defined as:

$$q = 1 - \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left(\cos(\psi_{jk}) + \frac{1}{3}\right)^2. \tag{2}$$

with ψ_{jk} the angle between the central water atom and neighbour water oxygen atoms j and k.

The translational tetrahedral order parameter, S_k, was also introduced by Chau and Hardwick, derived from an earlier order parameter defined by Kiselev *et al.*[11]. It measures the variance of the radial distances from a central water oxygen to its four neighbouring water atoms. It is defined as:

$$S_k = 1 - \frac{1}{3}\sum_{k=1}^{4}\frac{(r_k - \bar{r})^2}{4\bar{r}^{\,2}}. \tag{3}$$

where r_k is the radial distance from the central water oxygen to its kth neighbour and $\bar{r}$ is the mean of the four radial distances. The normalisation constants ensure that S_k increases with tetrahedral order, with a maximum value of 1 for a perfect tetrahedron. S_k has been previously shown to correlate more accurately with bulk water density than other order parameters, including q[13].

The Local Structure Index (LSI) was introduced by Shiratani and Sasai[14] and has previously been used to analysed protein hydration shells[15]. LSI quantifies the distance between the first and second hydration shells surrounding an individual water molecule, distinguishing molecules with well separated hydration shells from those in a disordered environment. We order the collection of radial oxygen-oxygen distances r_i for n neighbouring water atoms that are within 3.7 Å of the reference water oxygen so that $r_1 < r_2 < ...r_i < r_{i+1} < ...r_n < 3.7\mathring{A} < r_{n+1}$, the LSI is then given by:

$$LSI = \frac{1}{n}\sum_{i=1}^{n}(\Delta(i) - \overline{\Delta})^2. \tag{4}$$

where $\Delta(i) = r_{i+1} - r_i$ and $\overline{\Delta}$ is the arithmetic mean of $\Delta(i)$ over all neighbours within the cutoff distance of molecule i. A large LSI value corresponds to a highly structured tetrahedral coordination of water, with fewer molecules in interstitial positions.

The results of the order parameter analysis for the 260-protein dataset is shown in 26. The second hydration shell around proteins of the 260-protein dataset is on average more dense than the first hydration shell. The peak in density corresponding to the middle of the second hydration shell also coincides closely with peaks in the LSI order parameter, indicating that the individual water molecules within the second hydration shell have well-separated hydration shells as compared to the bulk water or first hydration shell. While q and s_k tends to increase with a greater gradient around the hydration shells, the values in both hydration shells are lower than that of bulk water, indicating that the neighbouring water molecules of water molecules within the hydration shells are radially and orientationally aligned less closely to the perfect tetrahedral alignment than for bulk water.

Supplemental Tables

Physical quantities	**Pearson's Correlation Coefficients (PCC)**	**PCC p-values**	**Spearmans's Correlation Coefficients (SCC)**	**SCC p-values**
Mass / 10^6 u	-0.078	0.274	-0.148	**0.036**
Mass (incl. buried water) / 10^6 u	-0.077	0.276	-0.150	**0.033**
Volume / 10^3 Å^3	-0.092	0.193	-0.174	**0.013**
Volume (incl. buried water) / 10^3 Å^3	-0.092	0.195	-0.173	**0.014**
Radius of gyration / Å	-0.132	0.061	-0.130	0.067
Num. waters	-0.039	0.583	-0.104	0.140
SASA / Å^2	-0.102	0.151	-0.180	**0.010**
Aspect Ratio	0.168	**0.017**	0.123	0.083
Density change / %	-0.060	0.400	-0.037	0.604
Charged % surface	-0.267	**0.000124**	-0.250	**0.000338**
Hydrophobic % surface	-0.290	**2.96e-05**	-0.254	**0.000279**
Other % surface	0.471	**1.70e-12**	0.445	**3.51e-11**
Acidic charged % surface	0.100	0.158	0.093	0.190
Basic charged % surface	-0.361	**1.40e-07**	-0.363	**1.17e-07**
Aliph. hydrophobic % surface	-0.290	**2.96e-05**	-0.254	**0.000279**
Arom. hydrophobic % surface	0.212	**0.00250**	0.147	**0.0378**
Aromatic % surface	0.358	**1.77e-07**	0.308	**8.93e-06**
Aliphatic % surface	-0.358	**1.77e-07**	-0.308	**8.93e-06**
Charged % interior	-0.0663	0.350	-0.0491	0.489
Hydrophobic % interior	**-0.696**	**1.98e-30**	**-0.672**	**8.39e-28**
Other % interior	**0.675**	**4.10e-28**	**0.625**	**3.92e-23**
Acidic charged % interior	0.205	**0.00353**	0.208	**0.00300**
Basic charged % interior	-0.207	**0.00315**	-0.184	**0.00876**
Aliph. hydrophobic % interior	**-0.715**	**1.04e-32**	**-0.689**	**1.14e-29**
Arom. hydrophobic % interior	0.0721	0.309	0.0762	0.283
Aromatic % interior	0.305	**1.10e-05**	0.329	**1.87e-06**
Aliphatic % interior	-0.305	**1.10e-05**	-0.329	**1.87e-06**
Overall Charge	-0.289	**3.12e-05**	-0.328	**1.99e-06**
Charged %	-0.152	**0.0308**	-0.142	**0.0446**
Hydrophobic %	**-0.738**	**6.81e-36**	**-0.706**	**1.23e-31**
Other %	**0.644**	**6.60e-25**	**0.588**	**4.67e-20**
Acidic charged %	0.113	0.111	0.104	0.143
Basic charged %	-0.317	**4.45e-06**	-0.298	**1.73e-05**
Aliph. hydrophobic %	**-0.741**	**2.91e-36**	**-0.698**	**1.10e-30**
Arom. hydrophobic %	0.0342	0.630	0.0352	0.620
Aromatic %	0.244	**0.000495**	0.261	**0.000182**
Aliphatic %	-0.244	**0.000495**	-0.261	**0.000182**
coil %	0.401	**3.80e-09**	0.345	**5.33e-07**
helix %	-0.467	**2.84e-12**	-0.485	**2.98e-13**
strand %	0.341	**7.55e-07**	0.346	**4.73e-07**

Table S I. ***All physical quantities tested for correlation with protein mass density, with the appropriate Pearson's and Spearman's correlation coefficients.*** *Correlations with p-values below the threshold value of 0.05 in bold, and the correlation coefficient and associated p-value are in bold. See also Figure S11.*

Amino acid prevalence	**Pearson's Correlation Coefficients (PCC)**	**PCC p-values**	**Spearmans's Correlation Coefficients (SCC)**	**SCC p-values**
VAL %	-0.326	**2.29E-06**	-0.261	**0.000186**
ARG %	-0.078	0.274	-0.0574	0.418
SER %	0.263	**0.000160**	0.267	**0.000126**
LEU %	**-0.585**	**8.15E-20**	**-0.593**	**1.80E-20**
ASN %	0.256	**0.000243**	0.224	**0.00138**
CYS %	**0.539**	**1.60E-16**	0.397	**5.21E-09**
THR %	0.0841	0.235	0.0680	0.338
ASP %	0.185	**0.00872**	0.170	**0.0158**
GLN %	-0.135	0.0559	-0.119	0.0931
LYS %	-0.271	**9.87E-05**	-0.313	**6.17E-06**
MET %	-0.189	**0.00726**	-0.163	**0.0205**
GLY %	0.221	**0.00165**	0.232	**0.000901**
PRO %	0.122	0.0833	0.0310	0.662
TYR %	0.387	**1.41E-08**	0.376	**3.70E-08**
GLU %	-0.0109	0.877	-0.00643	0.928
ALA %	-0.176	**0.0127**	-0.108	0.127
HIS %	-0.0346	0.626	-0.0566	0.425
PHE %	-0.128	0.0696	-0.111	0.116
ILE %	-0.310	**7.49E-06**	-0.314	**5.54E-06**
TRP %	0.251	**0.000326**	0.210	**0.00273**

Table S II. ***Amino acid prevalence tested for correlation with protein mass density, with the appropriate Pearson's and Spearman's correlation coefficients.*** *Correlations with p-values below the threshold value of 0.05 in bold, and the correlation coefficient and associated p-value are in bold. See also Figure S12.*

Amino acid classification	**3-letter amino acid code**
Aromatic	PHE, TRP, TYR, HIS
Aliphatic	VAL, ARG, SER, LEU, ASN, CYS, THR, ASP, GLN, LYS, MET, GLY, PRO, GLU, ALA, ILE
Hydrophobic	ALA, VAL, LEU, ILE, PHE, MET, TRP
Charged	LYS, ARG, ASP, GLU, HIS
Other	SER, ASN, CYS, THR, GLN, GLY, PRO, TYR
Acidic charged	ASP, GLU
Basic charged	ARG, HIS, LYS
Aliphatic hydrophobic	ALA, VAL, LEU, ILE, MET
Aromatic hydrophobic	PHE, TRP

Table S III. ***Categorical classification of amino acids***. *These categories were used as features to train random forests regressors (RFRs).*

Model training	**Adjusted R^2**	**PCC**
All protein conformations structural and sequence features	Training: 0.945 ± 0.001 Test: 0.496 ± 0.077	Training: 0.9723 ± 0.0004 Test: 0.75 ± 0.04
Protein mean values of structural features and sequences	Training: 0.941 ± 0.006 Test: 0.612 ± 0.059	Training: 0.977 ± 0.003 Test: 0.83 ± 0.03
Protein Sequences	Training: 0.914 ± 0.006 Test: 0.430 ± 0.059	Training: 0.969 ± 0.002 Test: 0.72 ± 0.04

Table S IV. ***Random forest regressor models performance using different training sets and features***

Supplemental Figures

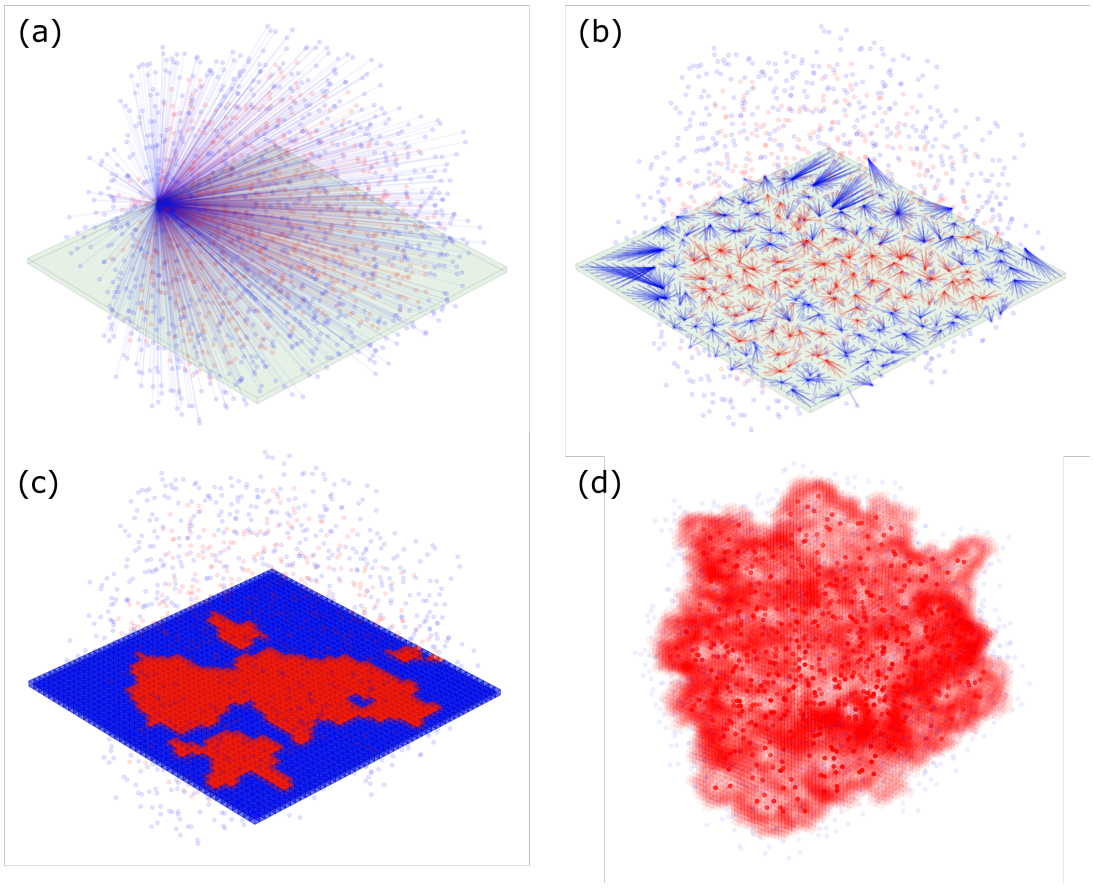

Figure S 1. **Schematic representation of the algorithm for protein density calculation.** *(a) For each layer (shown in green) of the grid (here of step size 1 Å) enclosing the protein, distances are calculated simultaneously from all grid points to all protein atoms (red points) and water atoms (blue points). The process is shown here for a single grid point, with vectors between the grid point and protein atoms represented by red lines and those between the grid point and water atoms by blue lines. At this point it is also possible to account for differing Van der Waals radii by subtracting the relevant radii from each distance. Then, as in (b), the distance to the nearest protein and water atoms respectively from each grid point are calculated (with whichever distance is shorter shown). Based on these shortest distances, voxels surrounding each grid point are defined as being part of the protein (red boxes) if they are closer to a protein atom than a water atom, as shown in (c). Blue boxes represent grid points that are closer to water atoms than protein atoms. (c) also demonstrates how a protein that appears approximately globular is in fact pierced deeply by water-accessible cavities. In (d), the final collection of voxels determined to represent the volume of the protein are shown.*

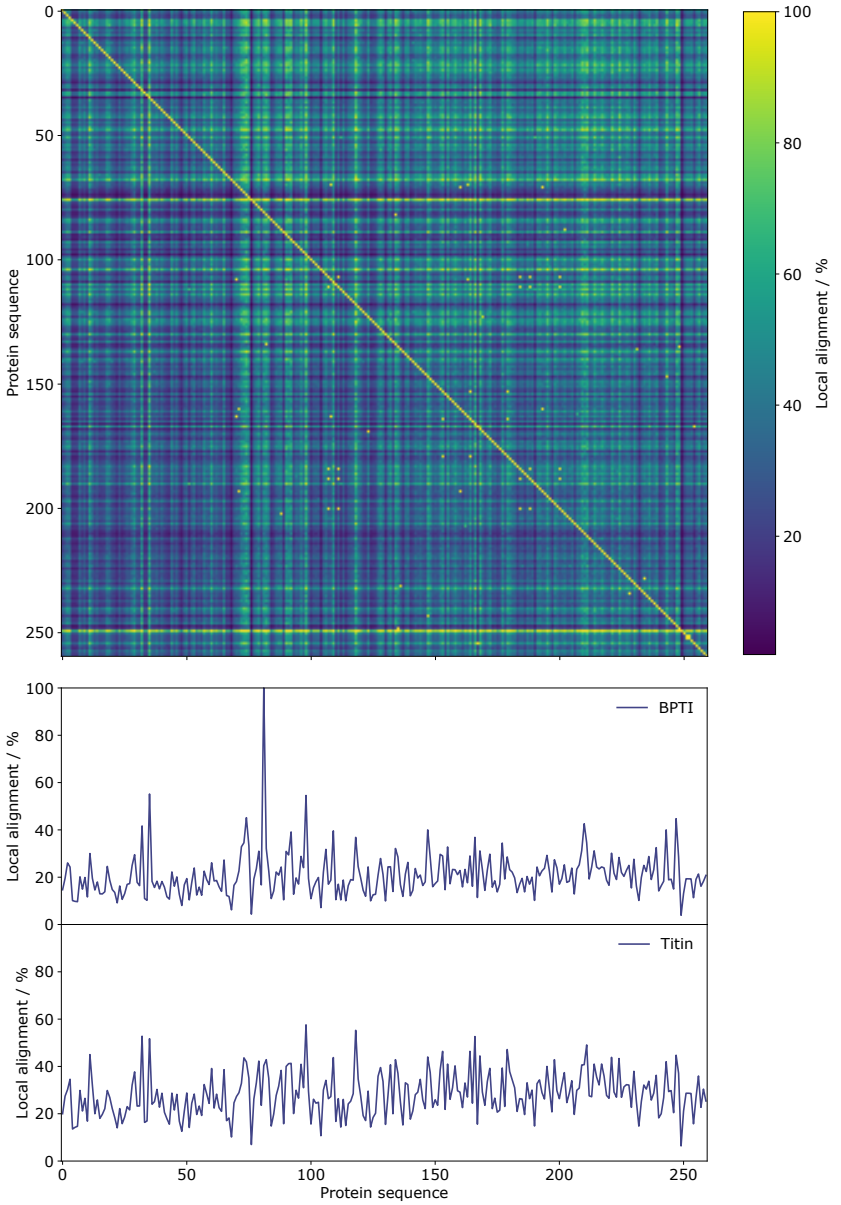

Figure S 2. **Local sequence alignment scores for the 260-protein dataset, and for titin and BPTI with the 260-protein dataset.** *Sequences in each column are locally aligned to the sequences in each row, with the alignment score (with 1 point for a correct residue match) converted to a percentage with respect to the length of the comparison sequence. The 100% present in the local alignment score for BPTI indicates that BPTI is present in the 260-protein dataset, whereas titin is not. Ignoring the diagonal, a mean local alignment percentage of 37 ± 0.1 % was found for the 260-protein dataset, compared to 28.3 ± 0.6 % for titin and the 260-protein dataset, and 20.8 ± 0.6 % for BPTI and the 260-protein dataset. Row 249 represents a large sequence from PDB 4FQI, which many other sequences can be well aligned to.*

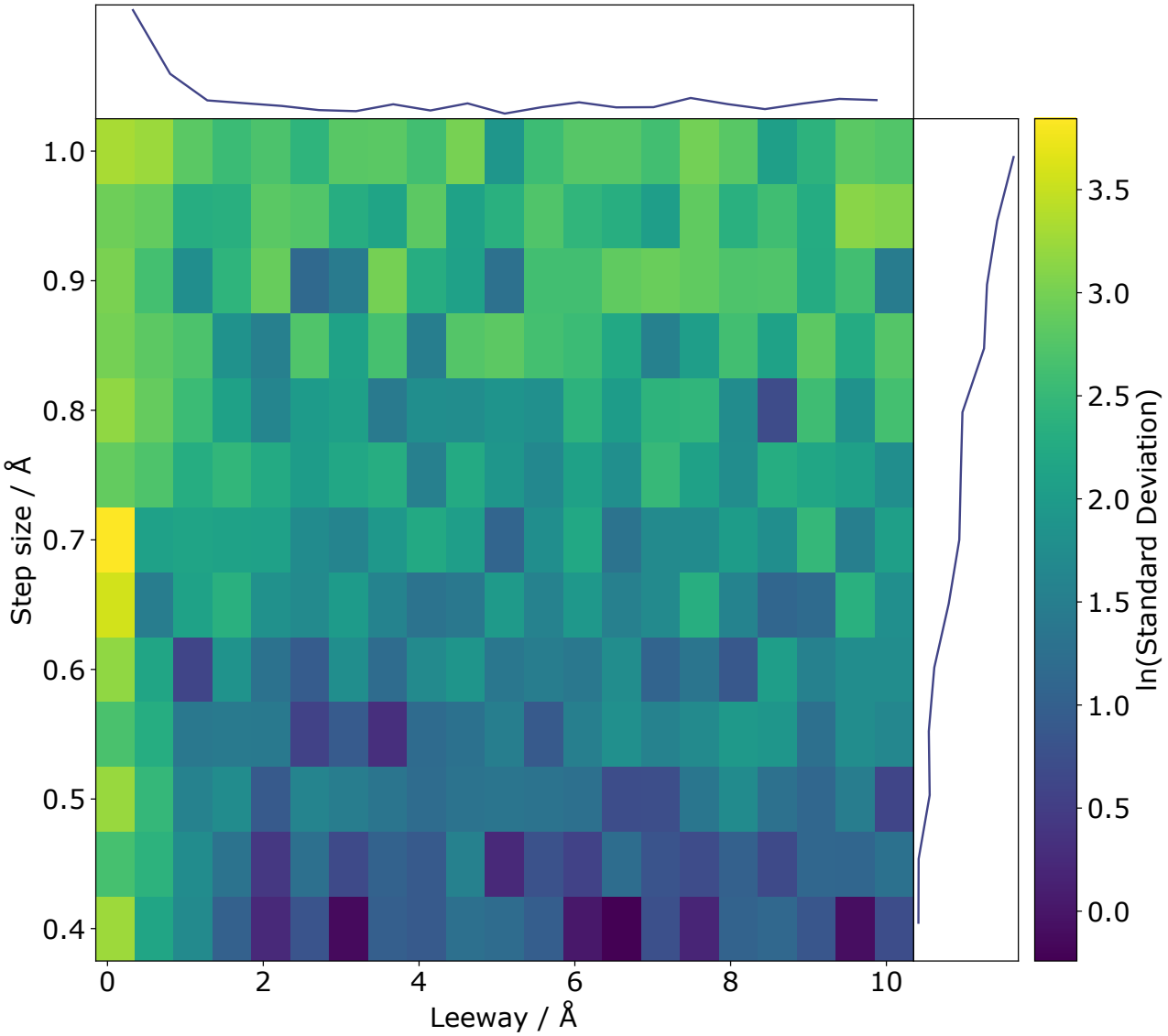

Figure S 3. **Benchmarking algorithm consistency depending on leeway and step size parameters.** *The natural logarithm of the standard deviations for a range of combinations of leeway and step size are shown, with 5 randomly rotated repeats for each combination of parameters for protein BPTI (PDB code: 5PTI). Inlaid line subplots represent the mean of the natural logarithm of the standard deviations along each step size value (right) and leeway value (top). Small leeway values below 1 Å lead to relatively high standard deviations, as waters are cut off around the extremes of a protein's position differently depending on how the protein is rotated relative to the grid. For higher leeway values, most waters should be included, reducing any leeway sensitivity. For larger leeway values, there is little correlation between leeway and standard deviation.*

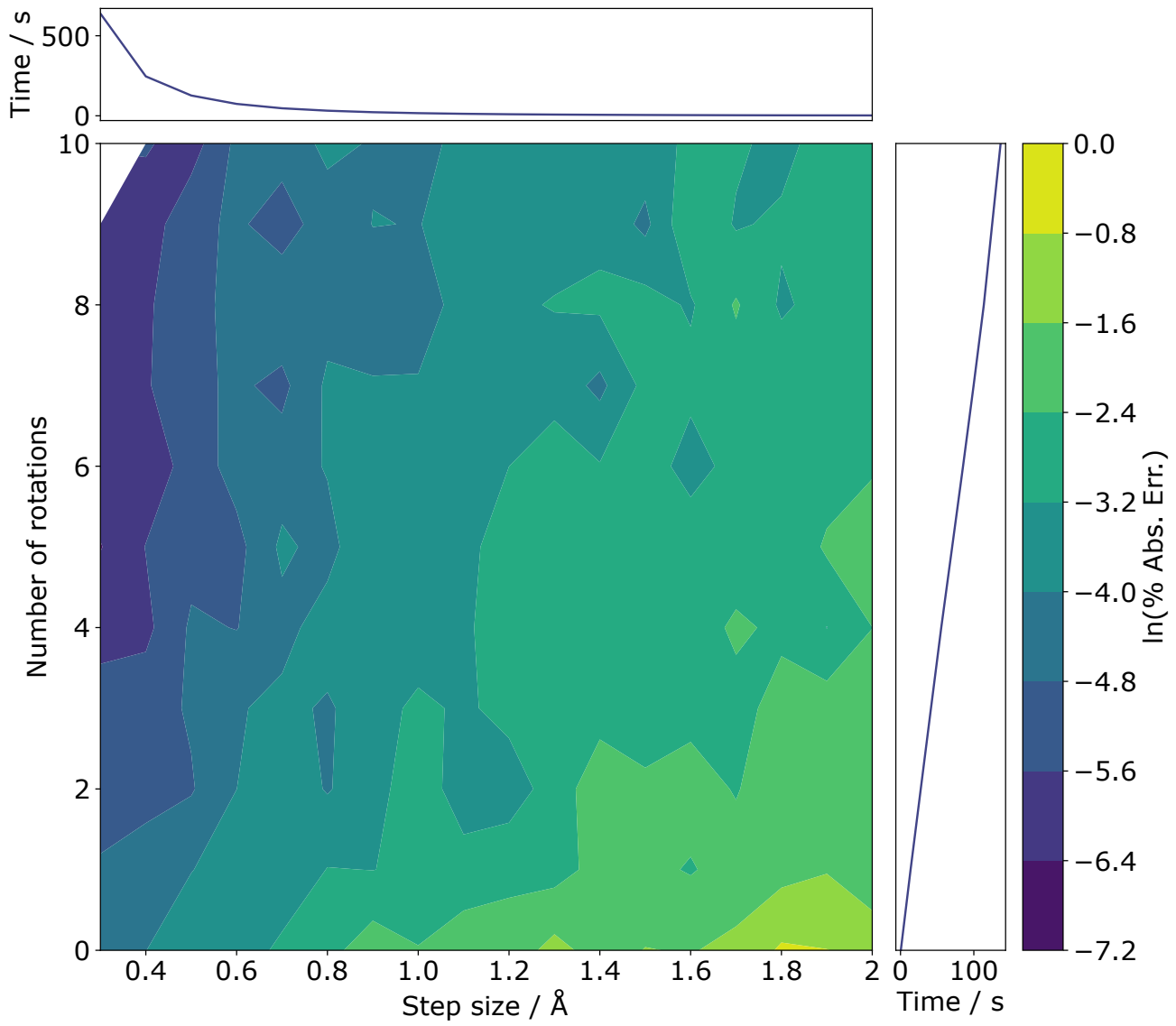

Figure S 4. **Benchmarking algorithm performance depending on rotations and step size.** *Results of parameter testing for 8 non-spherical protein structures from our dataset (PDBs: 3RVW, 2A9K, 1FQJ, 1IB1, 2VXT, 2O8V, 1FLE, and 1J2J), with sphericity ranging from 0.43 to 0.55. The CPU time averaged across each step size for all rotations (top) and for each number of rotations averaged across all step sizes (right) are shown. The calculation time increases linearly with the number of rotations, and cubically with decreasing step size. The percentage absolute error was calculated by comparing all volumes with the volume calculated using a step size of 0.3 Å and averaged over 10 rotations.*

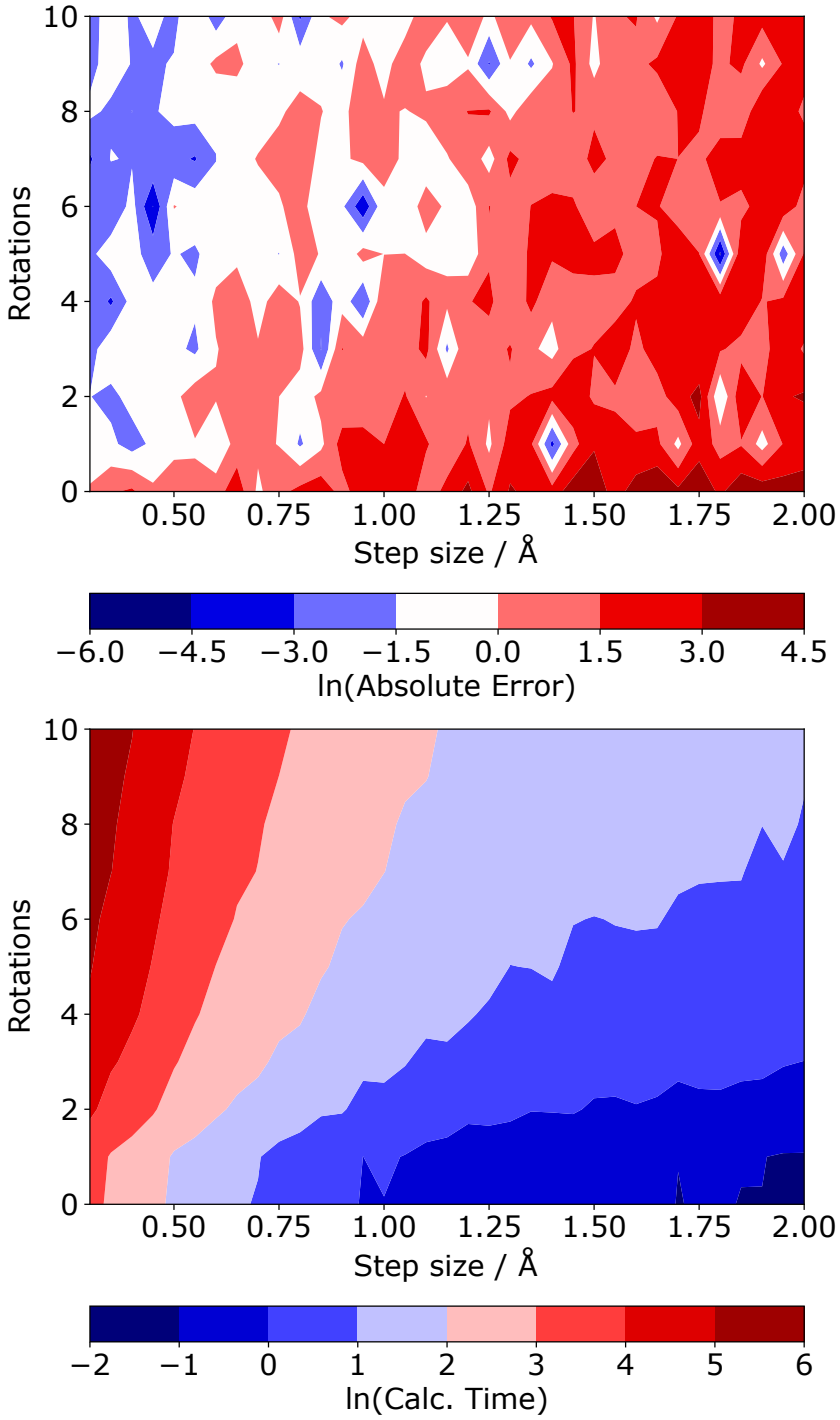

Figure S 5. **Benchmarking algorithm by varying rotation and step size parameters for BPTI.** *For BPTI a ground truth volume of 8279.4 ± 0.3 Å^3 was established by averaging the calculated volume over 1,000 random rotations, each averaged about 10 random axes with a 1.4 Å step size. This allowed the absolute error of volume calculations for BPTI with different parameters to be calculated. In general, combinations with smaller step sizes and more rotations are more accurate. However, these calculations ignored the effect of the atomic Van der Waals radii, which is likely why some parameter combinations with small step sizes and many rotations calculate lower than expected volumes.*

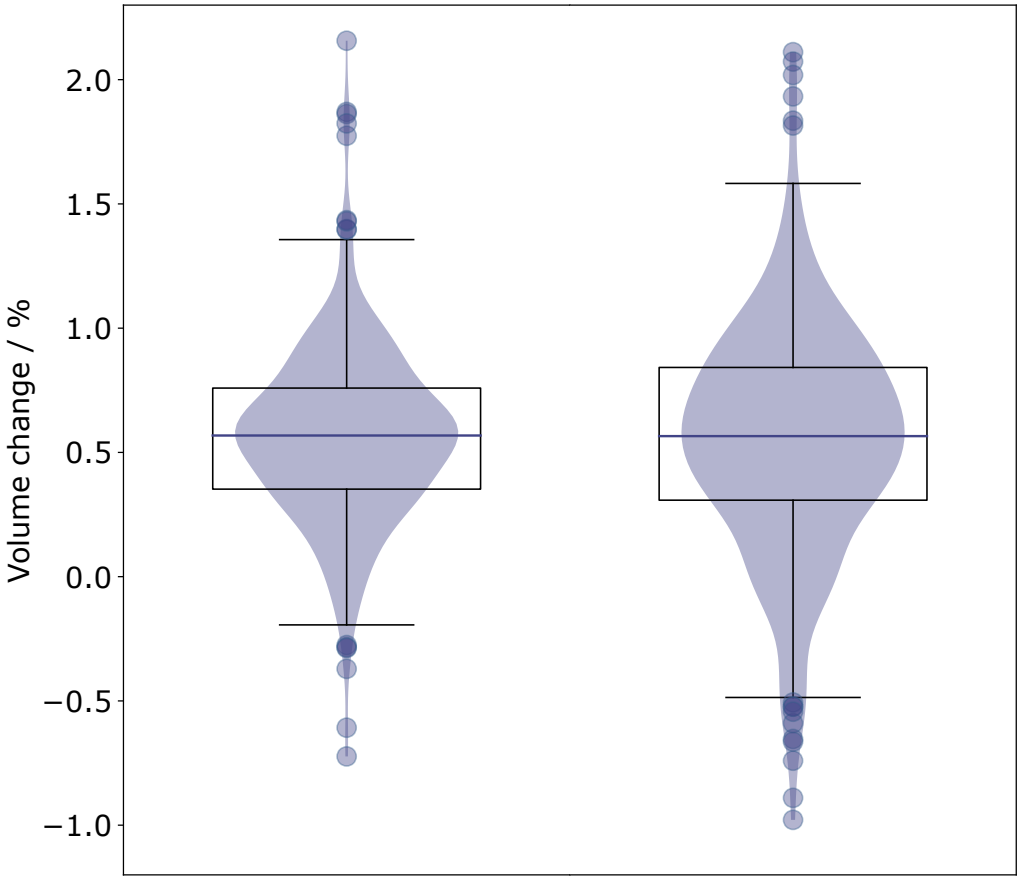

Figure S 6. **Volume change when the effects of van der Waals radii are included in calculations, with either standard or force field-specific radii values used.** *The change in calculated protein volume when the atomic Van der Waals radii are included in the consideration of whether a point in the 3-dimensional grid surrounding the protein is closer to a protein or water atom. Left, the atomic radii are taken from Bondi, 1964[16], giving a mean volume change of 0.57 ± 0.02 % (SD: 0.36). The upper and lower volume change percent outliers have densities that are respectively higher and lower than the non-outlier density (upper outliers, mean 1.33, standard deviation 0.03; non-outliers, mean 1.30, standard deviation 0.02; lower outliers, mean 1.26, standard deviation 0.01). Hence, including the effects of van der Waals radii reduces some of the highest calculated densities and increases some of the lowest calculated densities, narrowing the distribution. The radii for each atom can be extracted from the Amber ff14SB force field file (with hydroxyl and water hydrogens, given a radius of 0.0000 Å in Amber ff14SB corrected to a radius of 1.3870 Å, giving a mean volume change of 0.55 ± 0.02 % (SD: 0.47). Using these radii (right), the volumes are comparable to the results calculated both by ignoring the effects of the van der Waals radii and by naively including the effects by using a single radius value for each element (left).*

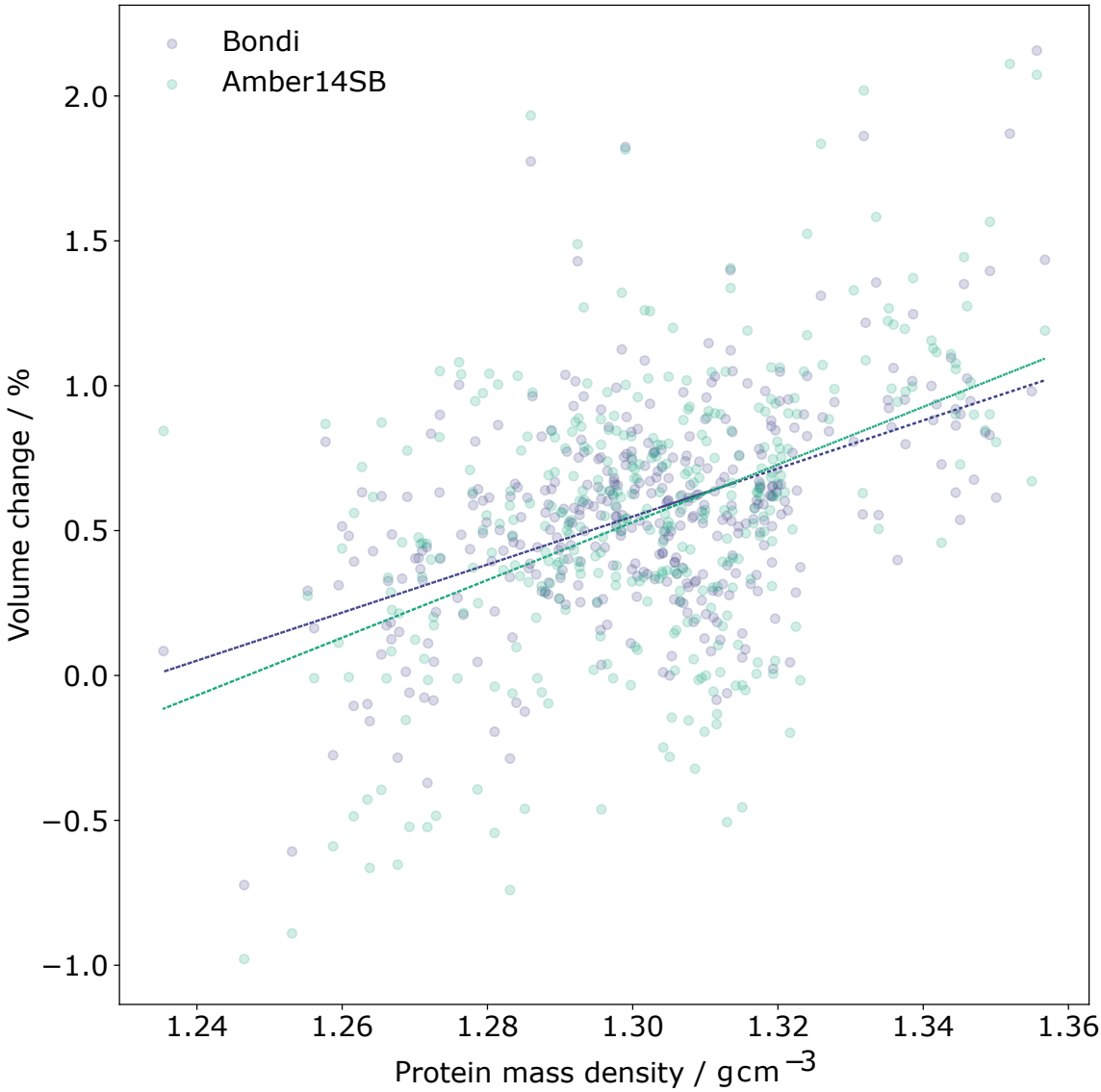

Figure S 7. **Volume change upon including atomic van der Waals radii in calculations is related to protein mass density.** *The volume change with including the effects of the van der Waals radii of atoms is positively correlated with the protein mass density, both while using Van der Waals radii from Bondi[16] or the Amber14SB force field[17]. Adjusted R-squared values of 0.242 (Bondi radii) and 0.198 (Amber14SB radii) are calculated for linear fits of the volume percentage change versus protein mass density, with gradients of 8.29 and 9.96 %(g cm^3)$^{-1}$. By increasing the volumes of proteins with the highest density by the largest amount, including the effects of the Van der Waals radii has the effect of narrowing the overall distribution of protein densities.*

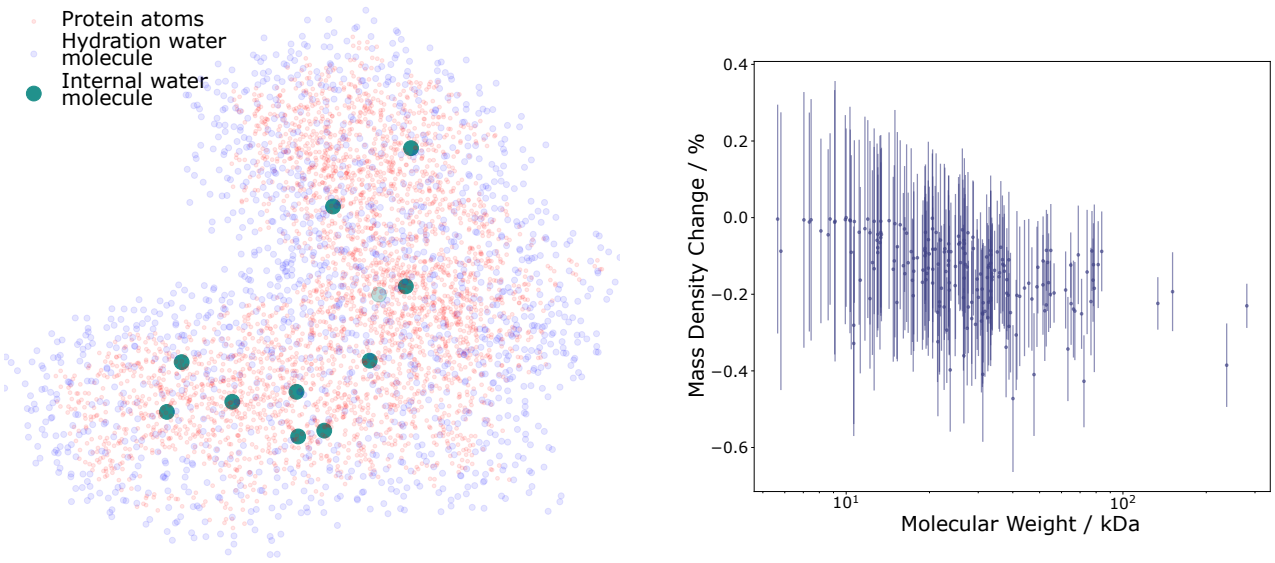

Figure S 8. ***Clustering of water molecules to identify internal water molecules, and the subsequent effect of internal water molecules on protein mass density.*** Left, clusters of water are defined using DBSCAN based on the positions of each water molecule's oxygen atom (eps = 3.0, min_cluster_size = 1) and plotted with the positions of protein atoms (red). Cluster 0, the largest cluster, contains hydration water surrounding the protein, while the next clusters contain waters that are buried inside the protein (and distance threshold of 3 Å is applied to ensure that usually positioned waters within the hydration shell or bulk are not incorrectly classified as buried waters). Water molecules within at least 6 Å of the protein must be retained prior to clustering, else the bulk water is unlikely to be identified as a single cluster. A mean of 8 internal waters is identified for the protein dataset, with a standard deviation of 12. Right, including the effects of these buried waters on the calculated protein mass densities (by incorporating their mass and considering them as part of the protein) has a minimal effect, with a mean mass density of 1.288 ± 0.002 gcm^{-3} (compared to 1.290 ± 0.002 gcm^{-3} if the effect of internal water molecules is neglected).

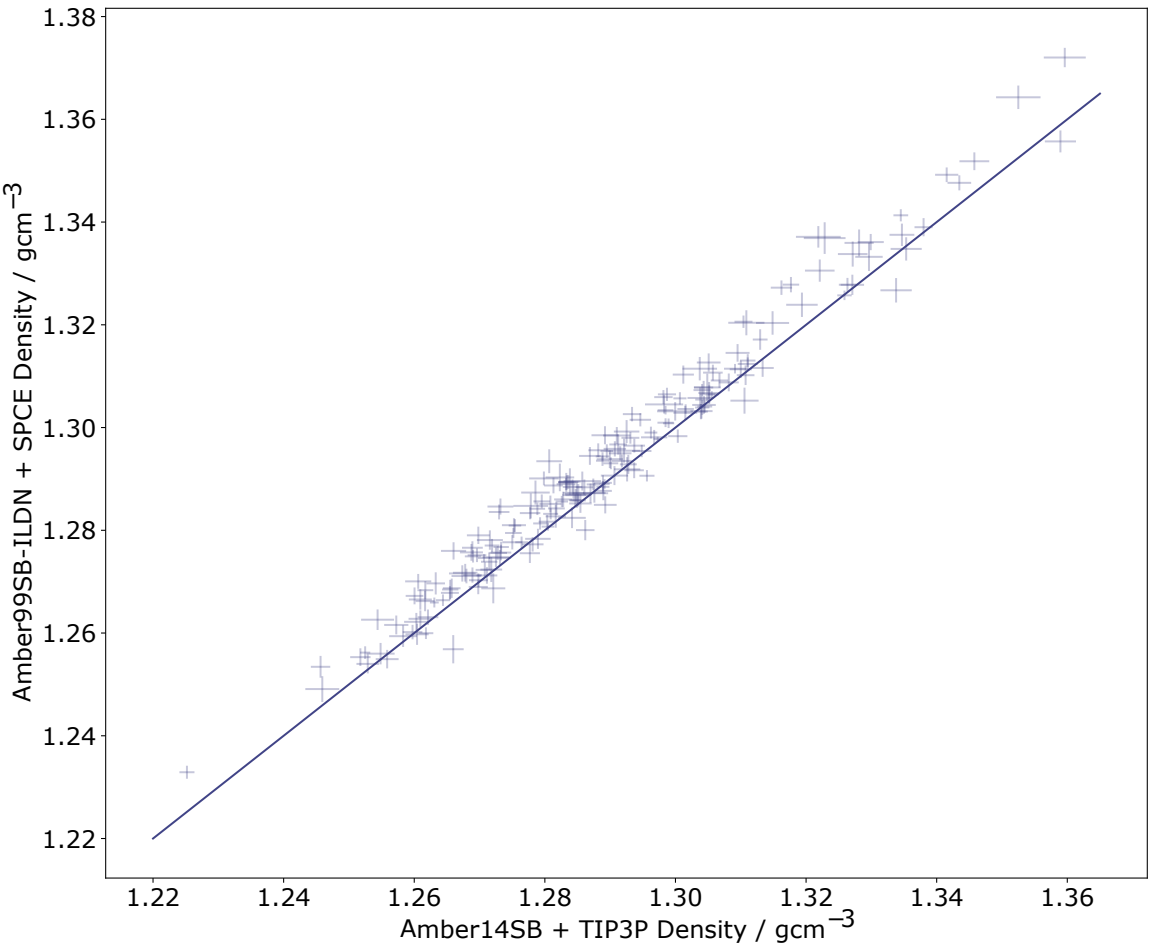

Figure S 9. **Comparison of calculated protein mass densities using two different force field and water model combinations.** *Protein mass densities calculated for the protein dataset using both the Amber99SB-ILDN force field and SPCE water model and the Amber14SB force field and TIP3P water model and averaged across 20 frames covering a 1 ns simulation, are plotted with standard error on the mean shown. The line of equality is plotted for comparison, with the protein mass density tending to lie slightly above the line, indicating that the densities calculated from simulations using the Amber99SB-ILDN force field with the SPCE water model are similar to those calculated from simulations using the Amber14SB force field and TIP3P water model, indicating that the density calculation procedure is robust to sensible changes to the force field and water model used. The TIP3P water model is known to underestimate the density of water[18]; larger spacing between molecules of water could explain the small discrepancy between the densities calculated using the SPCE and TIP3P water models.*

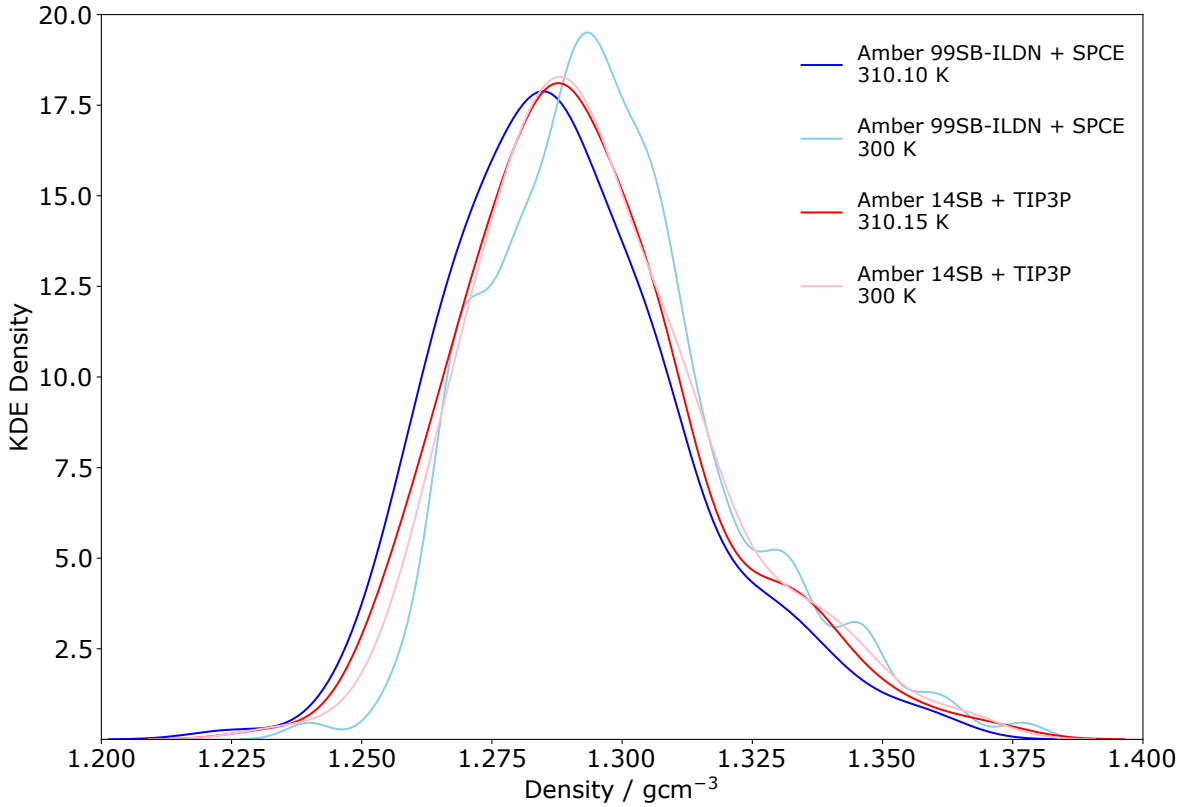

Figure S 10. **Gaussian Kernel Density Estimate (KDE) densities of the calculated densities of all protein-water conformations extracted from the molecular dynamics simulations of our protein dataset using different force field and water combination, at two temperatures.** *Simulating proteins using the Amber14SB force field and the TIP3P water model at different temperatures gives a subtle shift in the distribution of densities, with the mean density increasing from 1.293 $\pm$ 0.002 g cm^{-3} to 1.295 $\pm$ 0.002 g cm^{-3} (within the standard errors of the means) as temperature is reduced from 310.15 K to 300 K. However, the Amber99SB-ILDN and SPCE force field and water model combination seems more sensitive to the change in temperature, with the mean density increasing from 1.289 $\pm$ 0.002 g cm^{-3} to 1.298 $\pm$ 0.0004 g cm^{-3} with the same decrease in temperature.*

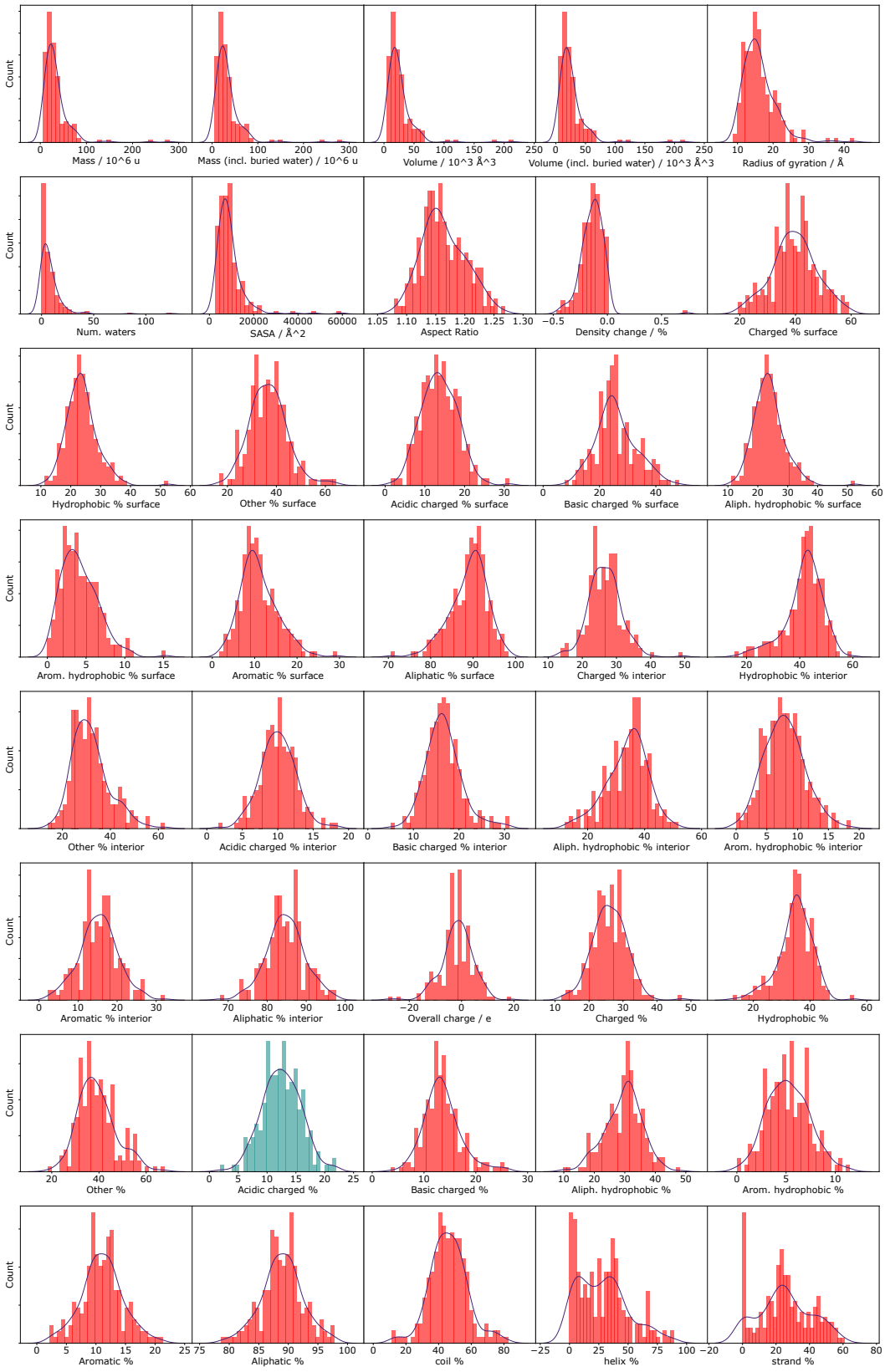

Figure S 11. **Normal distribution analysis for tested physical quantities.** *Histograms and gaussian kernel density plots for various physical quantities in the protein dataset. Plots of variables for which the Shapiro-Wilk Test for normality gives a p-value of greater than 0.5 are coloured in green, indicating that they may be normally distributed. Only one quantity, the percentage prevalence of acidic charged amino acid residues (in green), is approximately normally distributed.*

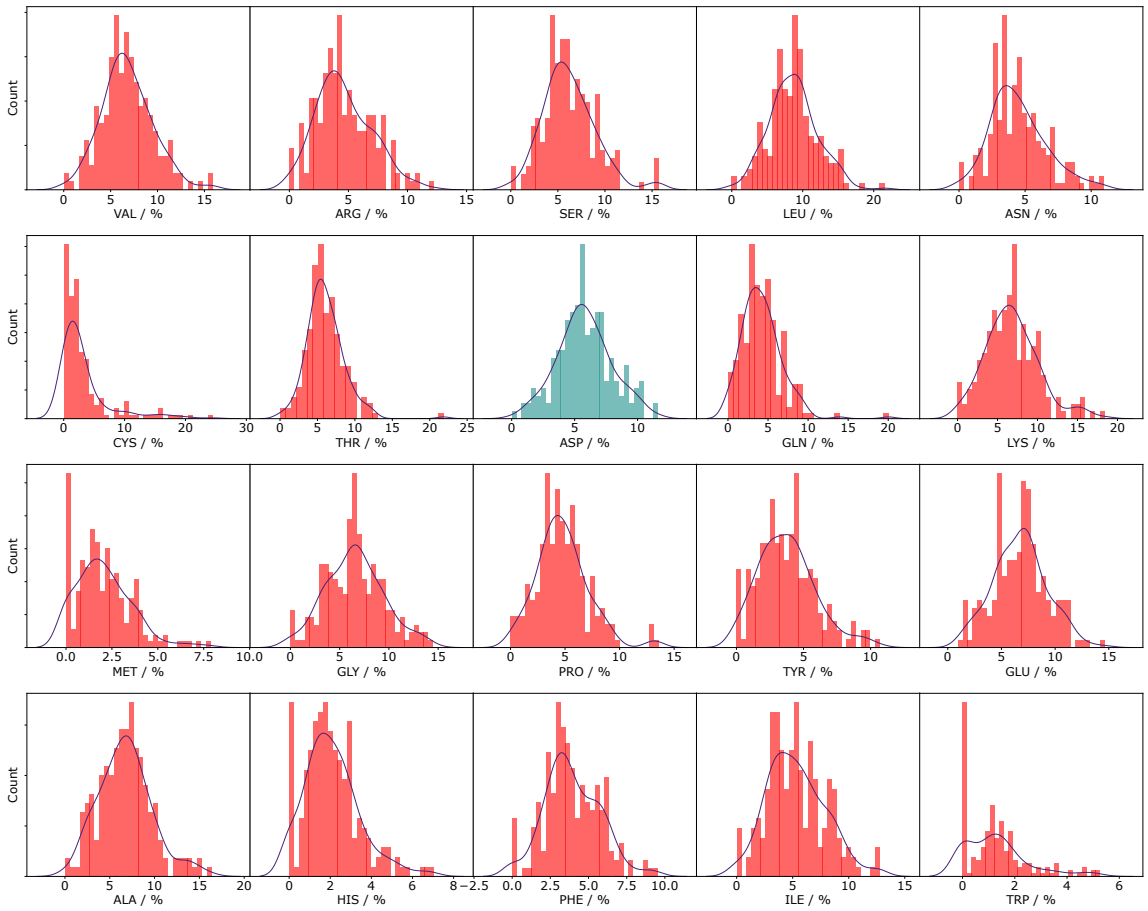

Figure S 12. **Normal distribution analysis for amino acid prevalence.** *Histograms and Gaussian kernel density plots for the distributions of the percentage prevalence of each standard amino acid for the protein dataset. Plots of variables for which the Shapiro-Wilk Test for normality gives a p-value of greater than 0.5 are coloured in green, indicating that they may be normally distributed. Only the percentage prevalence of Aspartic acid (in green) is approximately normally distributed.*

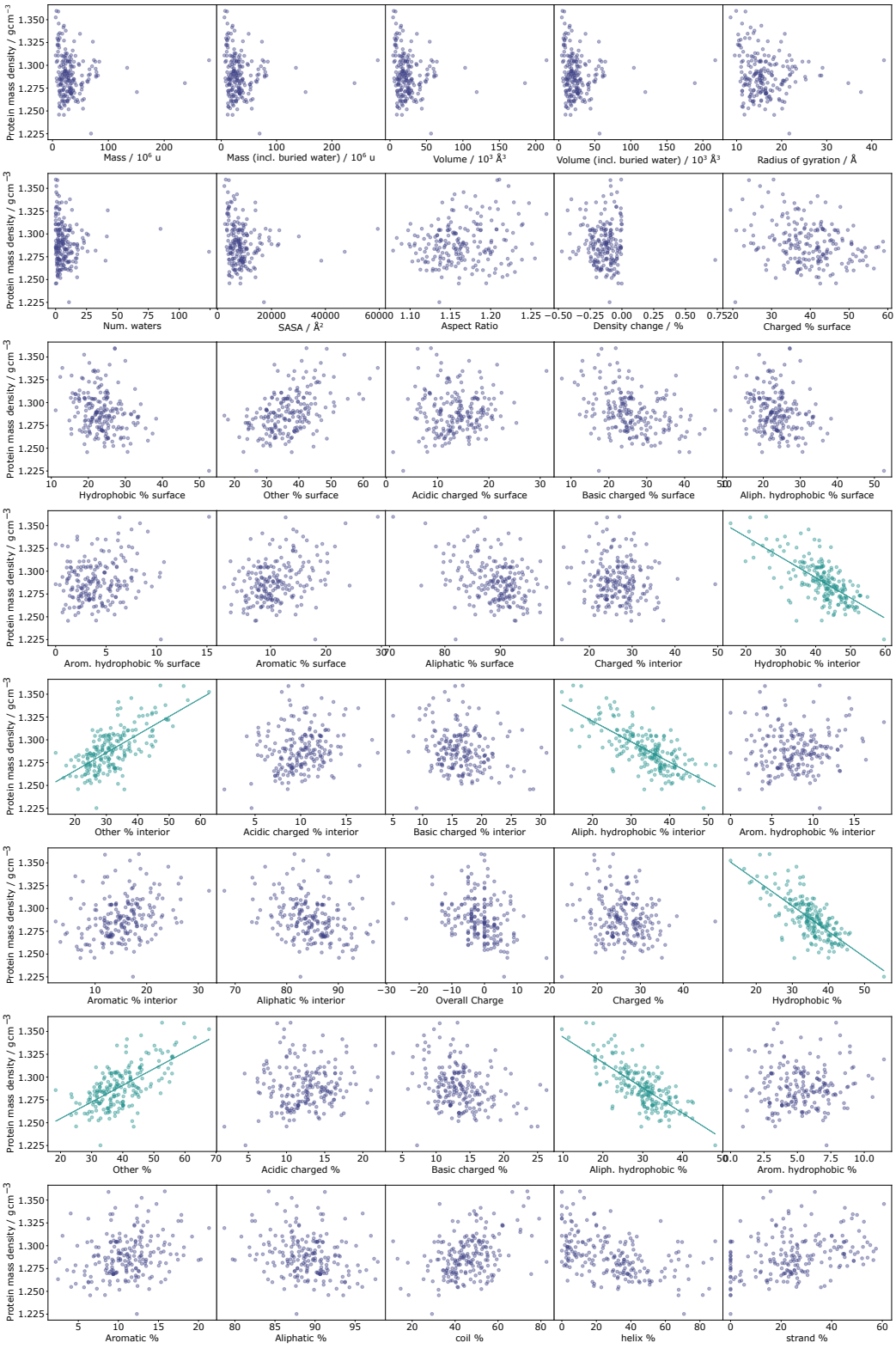

Figure S 13. **Spearman correlations between protein mass density and tested physical quantities.** *Protein mass density plotted against a range of physical parameters for the protein dataset. Plots in green indicate that a Spearman's correlation coefficient for correlation between these variables with absolute value greater than 0.5 and a p-value less than 0.05, implying a significant correlation. Such correlations are found between the protein mass density and several quantities representing the amino acid composition of the protein, in particular relating to the prevalence of hydrophobic residues. See also Table S1.*

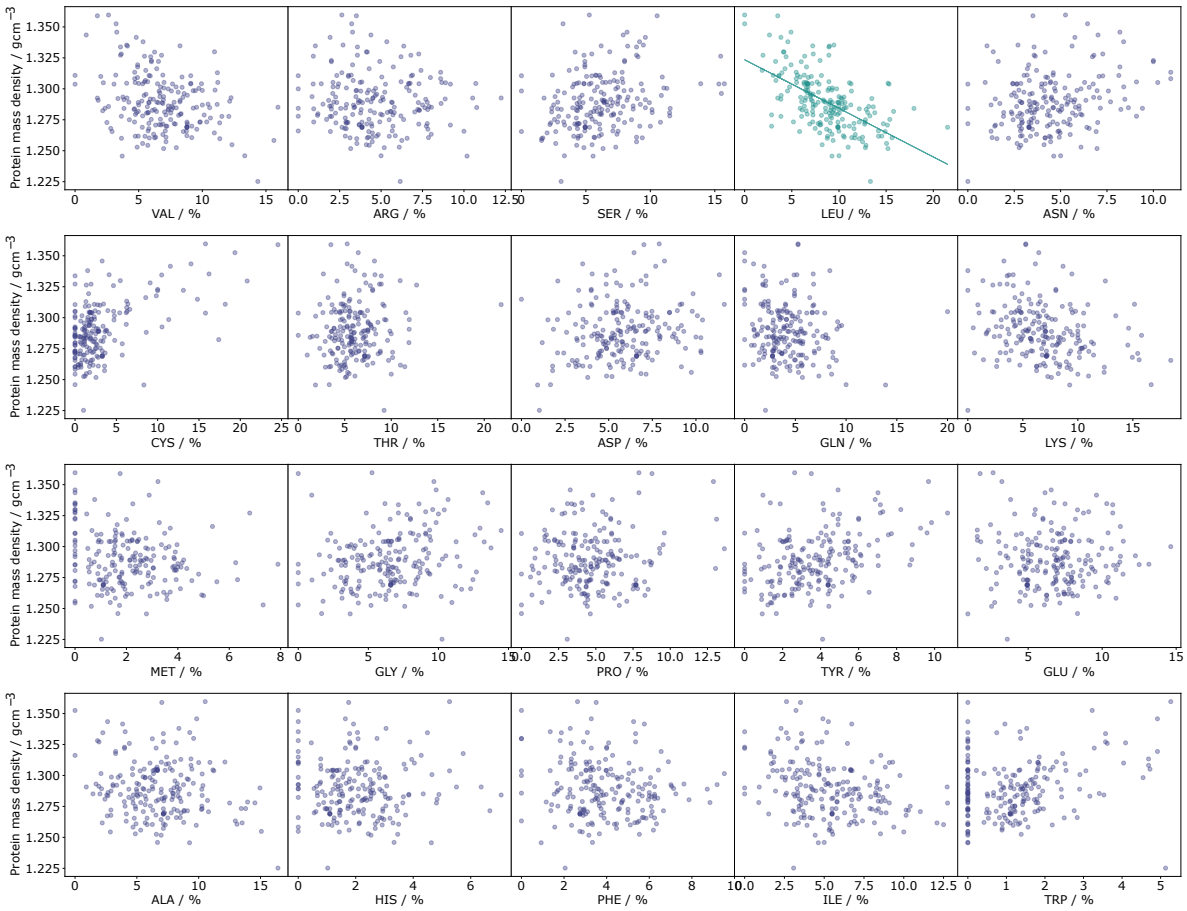

Figure S 14. **Spearman correlations between protein mass density and amino acid prevalence.** *Protein mass density and amino acid residue prevalence for the protein dataset. Only the relationship between protein mass density and the prevalence of Leucine has a Spearman's Correlation Coefficient (SCC) with absolute value greater than 0.5 (SCC: -0.5929) and a p-value of less than 0.05 (p-value: 1.80E-20). See also Table S2.*

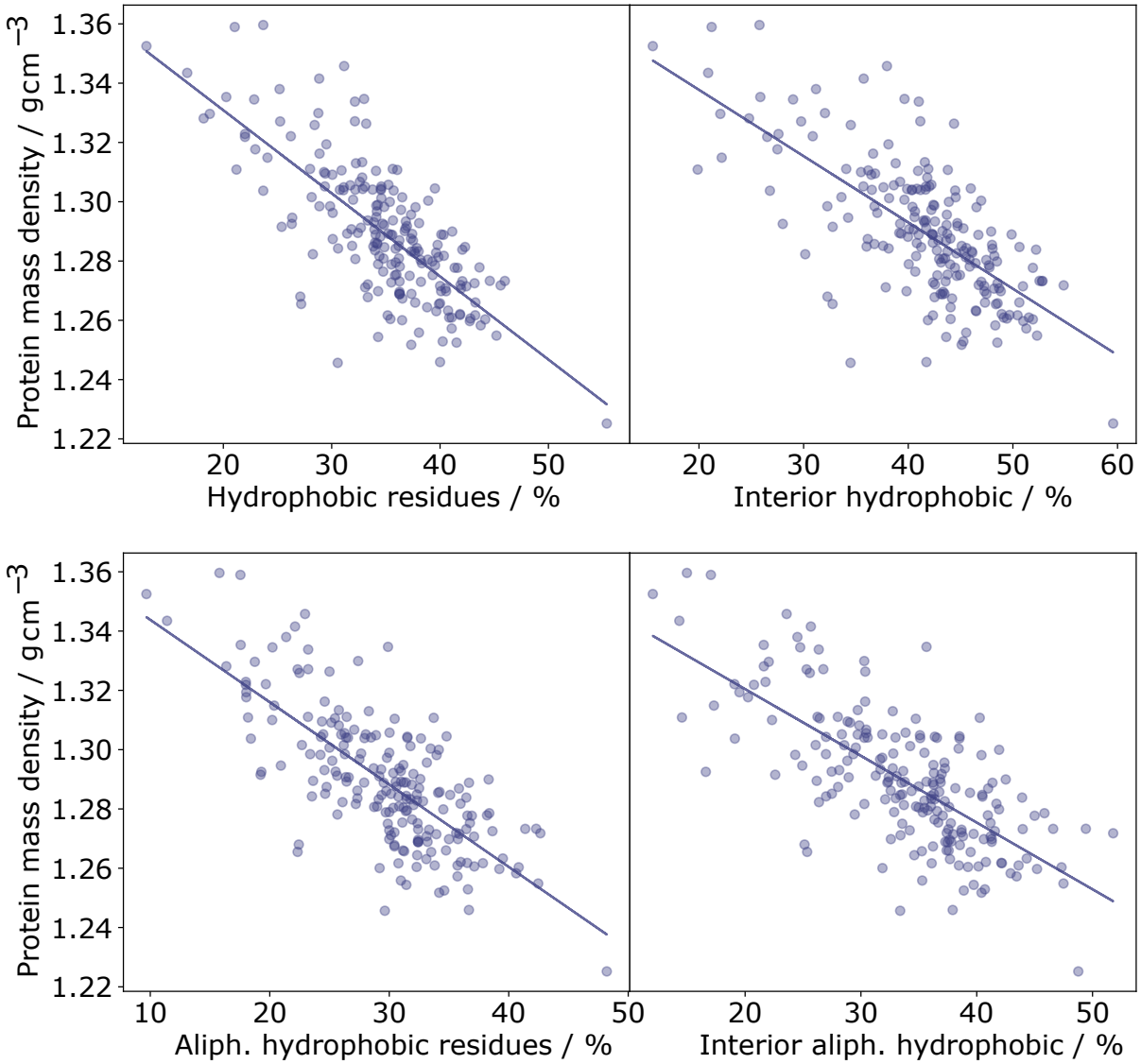

Figure S 15. **The protein mass density is well-correlated with the prevalence of hydrophobic residues.** *A Spearman's correlation coefficient of -0.7059 with a p-value of 1.23E-31 was found for the correlation between the protein mass density and the percentage of hydrophobic residues in the protein. A similar SCC of -0.6980 (p-value: 1.10E-30) was calculated for the correlation between the protein mass density of the percentage of aliphatic hydrophobic residues. The correlation seems to be predominantly explained by the interior residues, with an SCC of -0.6724 (p-value: 8.39E-28) between protein mass density and the prevalence of hydrophobic residues in the protein interior, and an SCC of -0.689286 (p-value: 1.14E-29) between the protein mass density and the prevalence of aliphatic hydrophobic residues in the protein interior.*

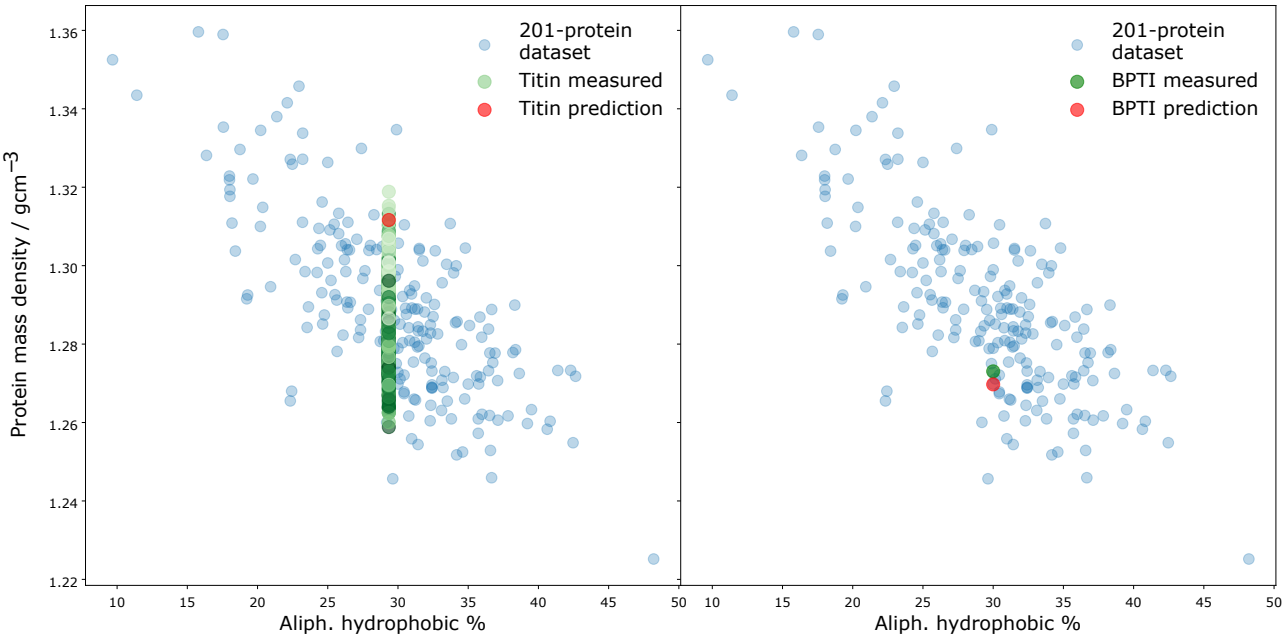

Figure S 16. **Comparison between measured and random forest predicted densities for the protein dataset, titin, and BPTI.** *The measured protein mass densities for the protein dataset are shown in comparison to the measured densities for titin (left) and BPTI (right). Measured densities for titin for all 873 energy minimised and equilibrated conformations between 1 ns and 6.8 ns are shown in green, with darker green indicating later time. The sequence-based RF (seq-RF) prediction for titin's density is close to the density measured for titin early in the pulling simulation. Later conformations of titin have reduced densities, but still lie within the density distribution of the protein dataset. A crystal structure of BPTI (PDB: 5PTI) is solvated and energy minimised before we calculate the density (1.273 gcm^3), which is close to the seq-RF prediction (1.270 gcm^3), and also lies within the distribution of the protein dataset.*

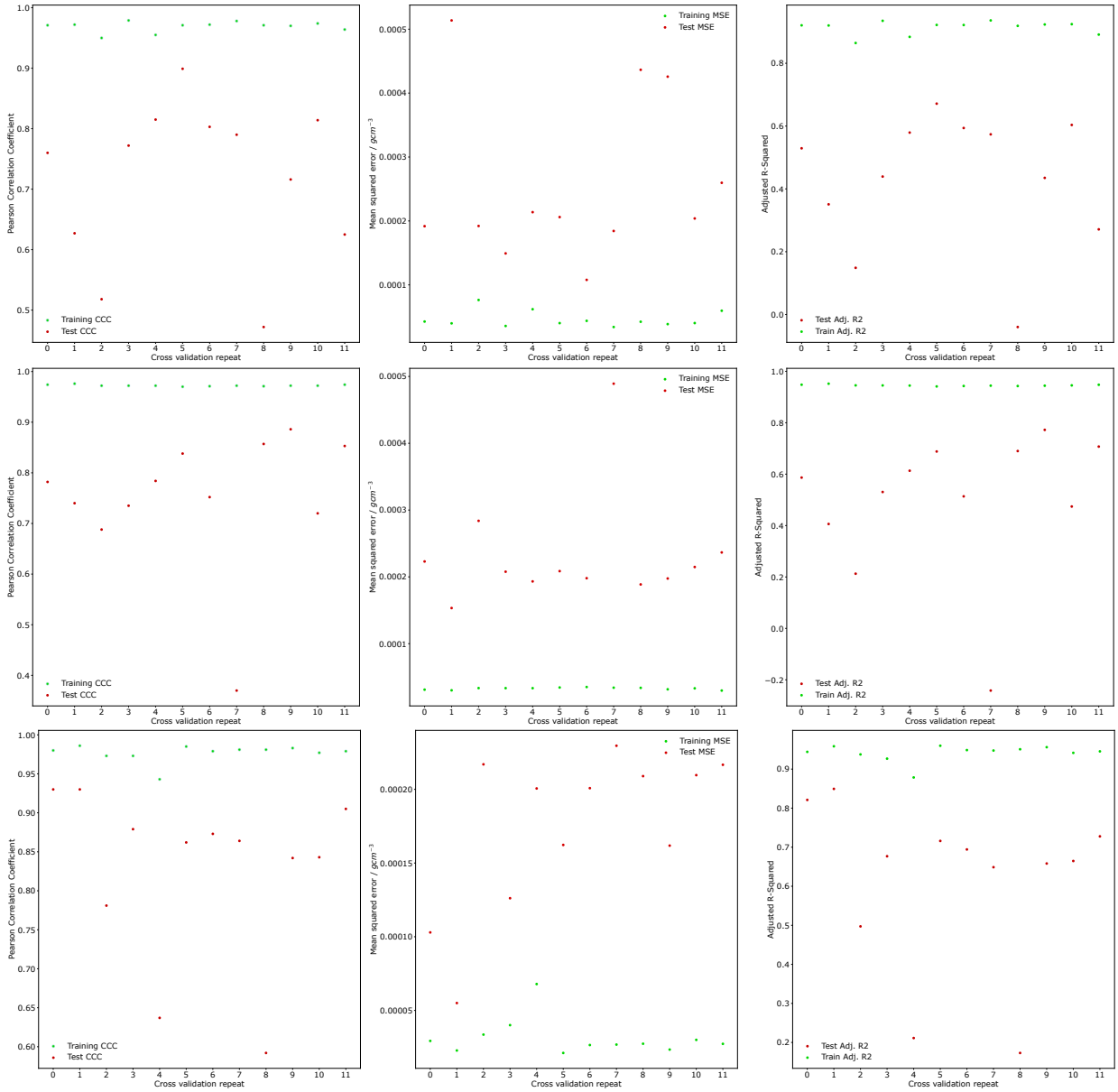

Figure S 17. **Evaluating the fit of the three random forest regressor models.** *The top row gives results for the random forest regressor model trained to predict protein densities on only protein sequences. The middle row gives the results for the random forest regressor model trained using structural feature values of all 5460 structures from the 260 protein dataset. The bottom row shows results for the random forest regressor trained using the mean structural feature values from the 5460 structures for each protein in the 260 protein dataset.*

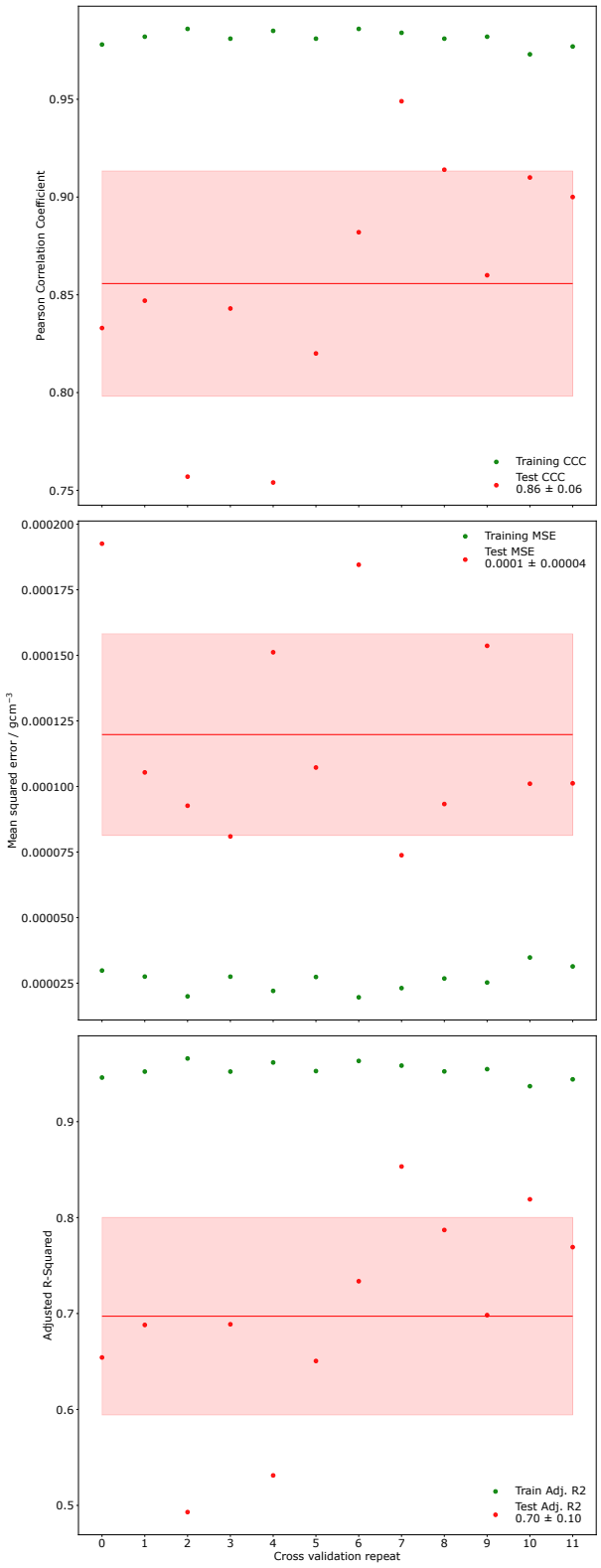

Figure S 18. **Evaluating the fit of the 20-feature random forest regressor.** *Evaluation of the performance of the random forest regressor model trained to predict protein densities for the 260-protein dataset using 20 features, with 12 repeats and grid optimisation for each repeat, gives a PCC for the predicted and computed densities for the test sets of 0.86 ± 0.06, indicating a good fit between the RF-predicted and computationally predicted protein densities. The mean for each statistic is shown as a red horizontal line, with the shaded bar representing the standard deviation.*

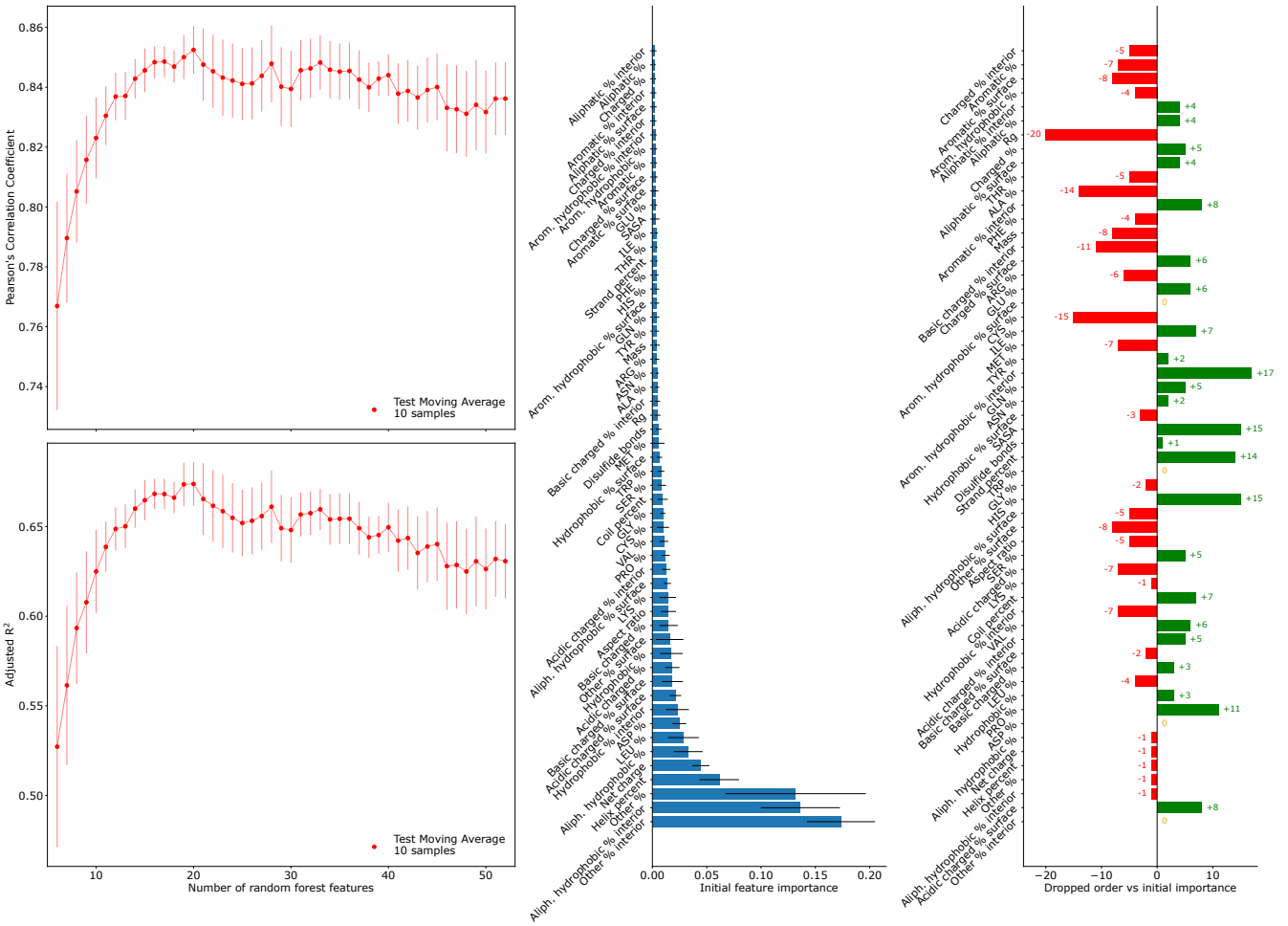

Figure S 19. **Evaluating feature importances via iterative feature removal.** *The relative importances of each feature of the full feature set (56 features) is assessed. In the scatter plots on the left the quality of the fit provided between the random forest-predicted and experimental densities are assessed via the Pearson's Correlation Coefficient (top) and the Adjusted-R^2 (bottom) for 10 grid-optimised random forest repeats, training initially on all features with the lowest feature by mean permutation importance removed successively until only the most important feature remains. As more features are added to the random forest regressors their predictive accuracy initially improves, but after around 20 features the increased complexity and potential overfitting worsens the performance of the random forest regressors (as well as increasing computation time). The middle bar plot shows the initial permutation feature importance for all 56 features, with standard deviation error bars. The bar plot on the right compares the position at which a feature is excluded to the relative initial permutation importance of the feature for a random forest trained on all 56 features, showing that there is a correlation between the initial importances and the order or exclusion, and that the variability of feature importance is inversely correlated to the importance itself.*

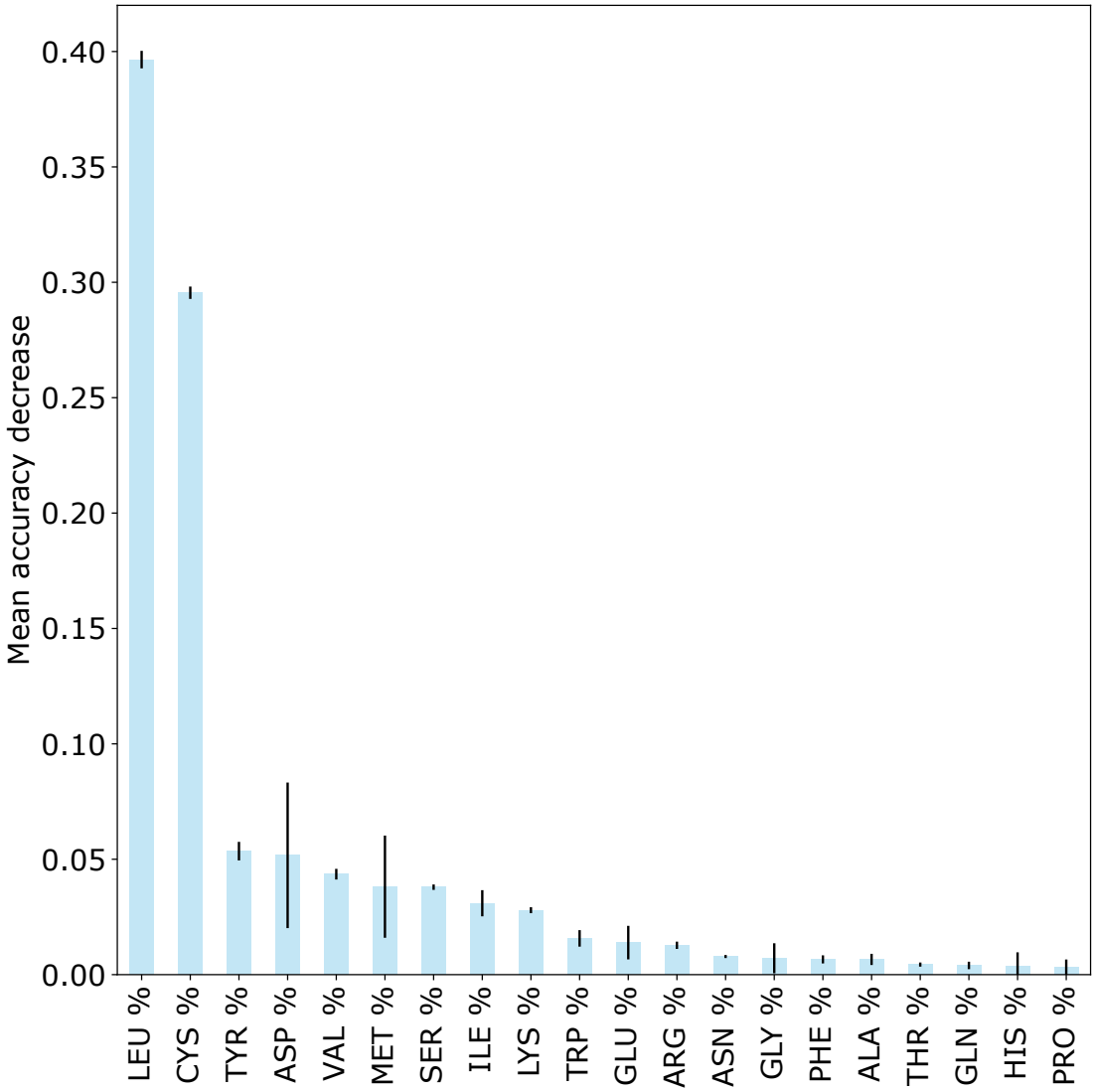

Figure S 20. **The importances of the prevalences of each amino acid residue to the density prediction random forest regressor.** *There is no correlation between the importance of a residue and its mass, volume, or density. However, there is a weak correlation (PCC: 0.45) between importance and residue hydrophobicity (as defined by the Kyte-Doolittle scale).*

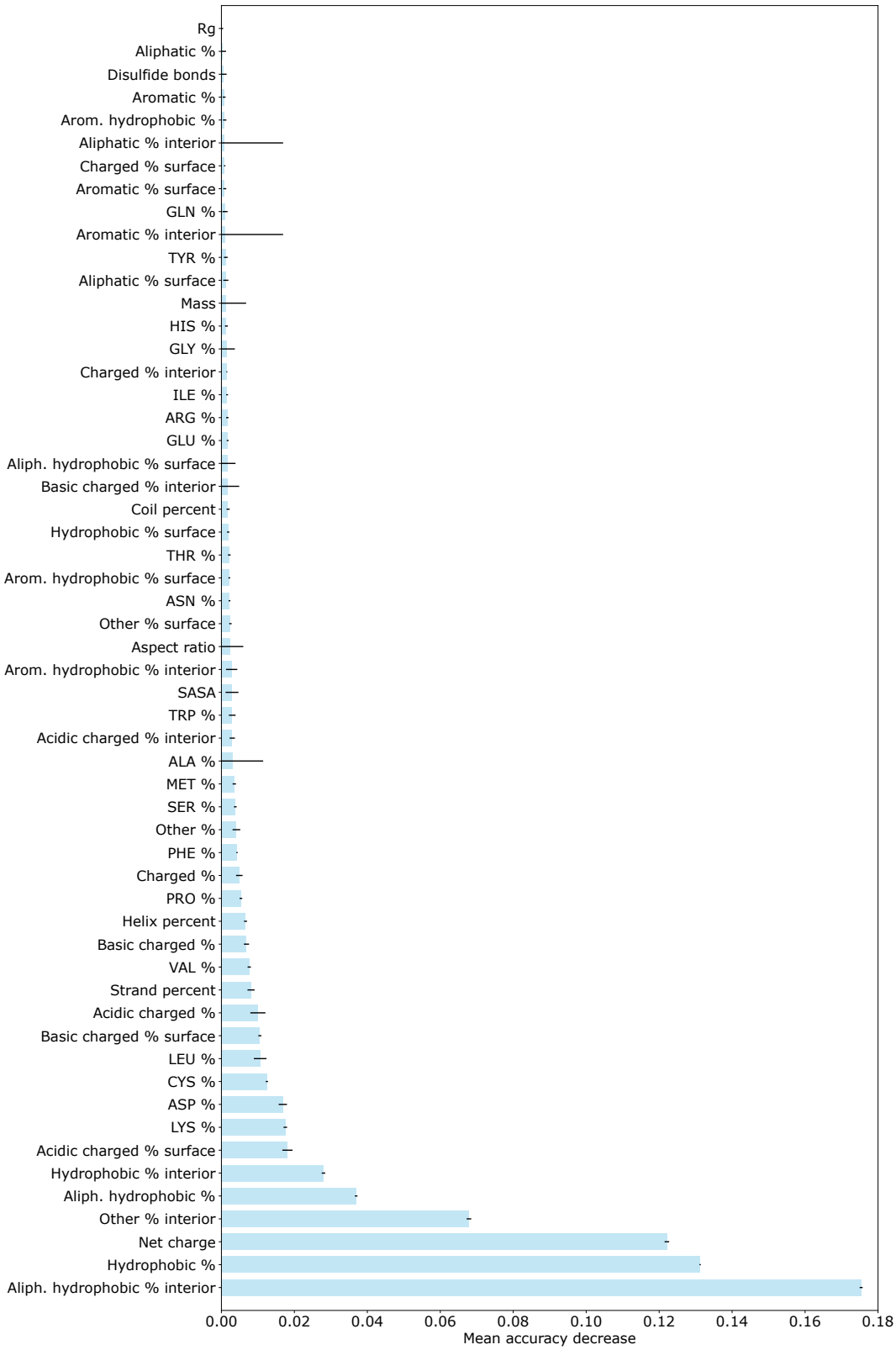

Figure S 21. **The full feature importances for all sequence- and structure-based features, trained on the sequences and mean structural characteristics of 4221 structures of the protein dataset.** *The percentage of atoms in the interior of the protein which are part of aliphatic hydrophobic residues (Alanine, Valine, Leucine, Isoleucine, and Methionine) is the most important feature to the random forest regressor when predicting protein densities, with the overall percentage of hydrophobic residues of the protein second. The net charge of the protein, the third most important feature, is calculated naively by considering all Glutamic acid and Aspartic acid residues to have a charge of -1 and all Arginine and Lysine residues to have a charge of +1. Out of the 56 features the random forest regressor is trained on, only 12 features have an importance of greater than 0.01.*

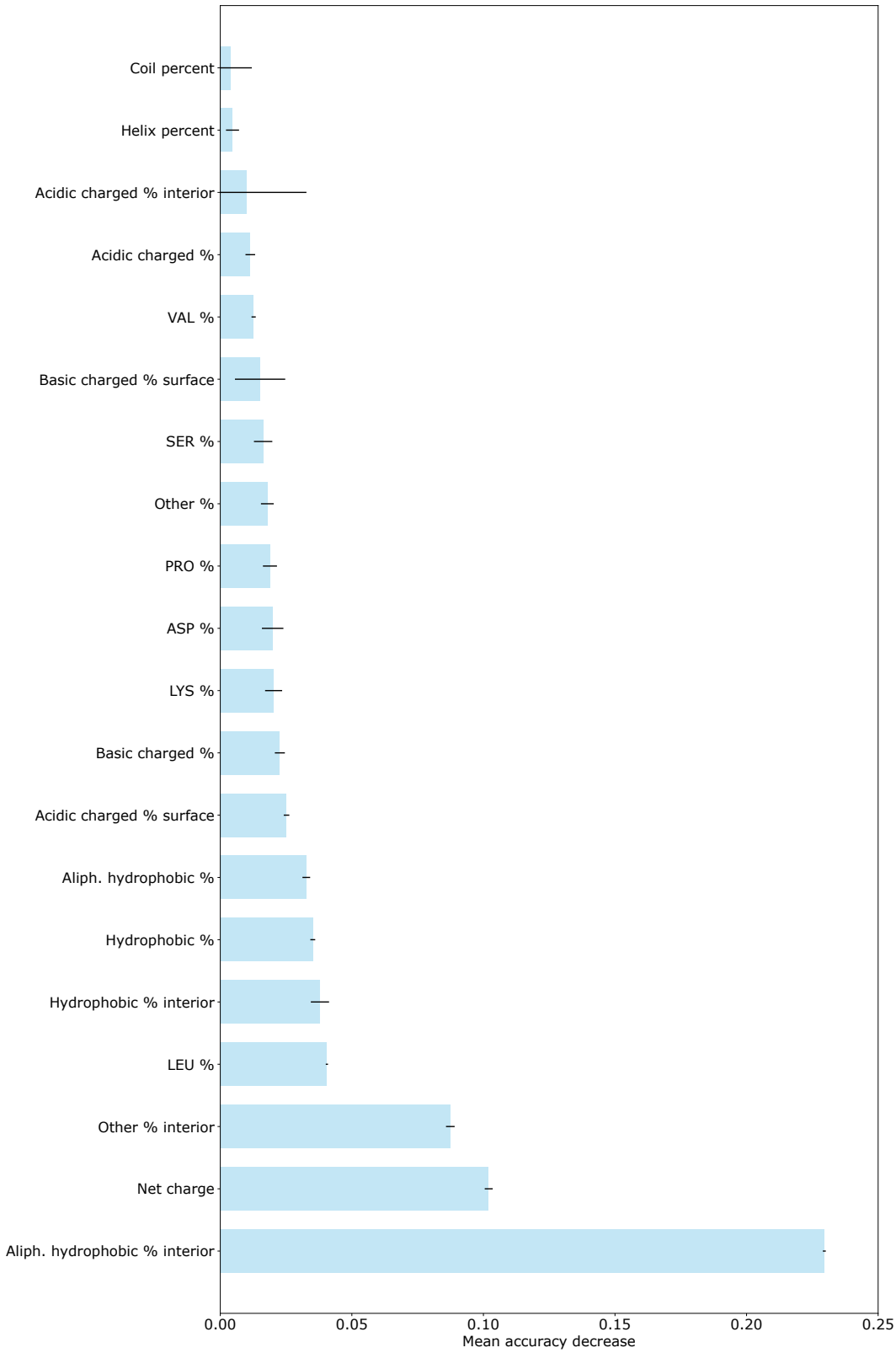

Figure S 22. **Struct-RF Regressor feature importances.** *The feature importances of the final Struct-RF Regressor, based on the 20 most important features selected from the ablation study (See Fig S19). The 20 features are chosen as they represent a peak in the accuracy of the RFRs generated and optimised with different sets of features, with the extra complexity of more features which are less relevant to our target characteristic (density) reducing performance.*

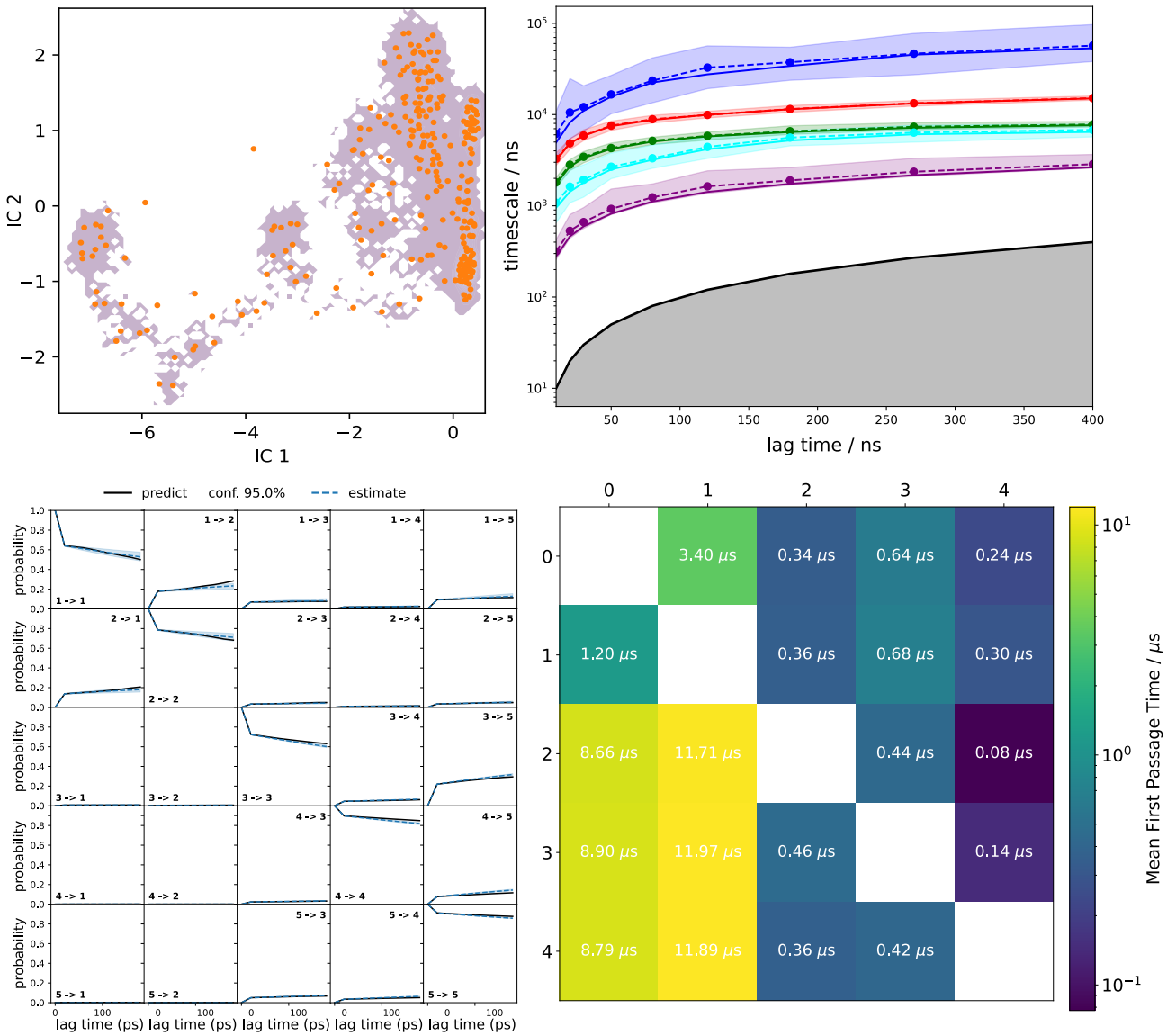

Figure S 23. **Markov State Model validation.** *Top left shows a simple apparent free energy landscape of the 1 ms BPTI simulation, with regions of high sampling along the first two independent components shaded and the 300 κ-means clusters overlaid in orange. The clusters are well-distributed in this low-dimensional TICA subspace. Top right shows the implied timescales (ITS) for maximum likelihood MSM in solid lines, with the dashed lines given by the sample means of samples generated by the Bayesian MSM, with confidence intervals containing 95% of the Bayesian MSM samples shown as shaded regions. At the chosen lag time of 20 steps (200 ns) the implied timescales have converged, within error. Bottom left shows the results of the Chapman-Kolmogorov test for this system with a lag time of 200 ns and 5 states chosen. Assuming 5 metastable states passes the Chapman-Kolmogorov test. Bottom right shows the transition matrix for the final MSM, also reporting on the mean first passage times for each transition.*

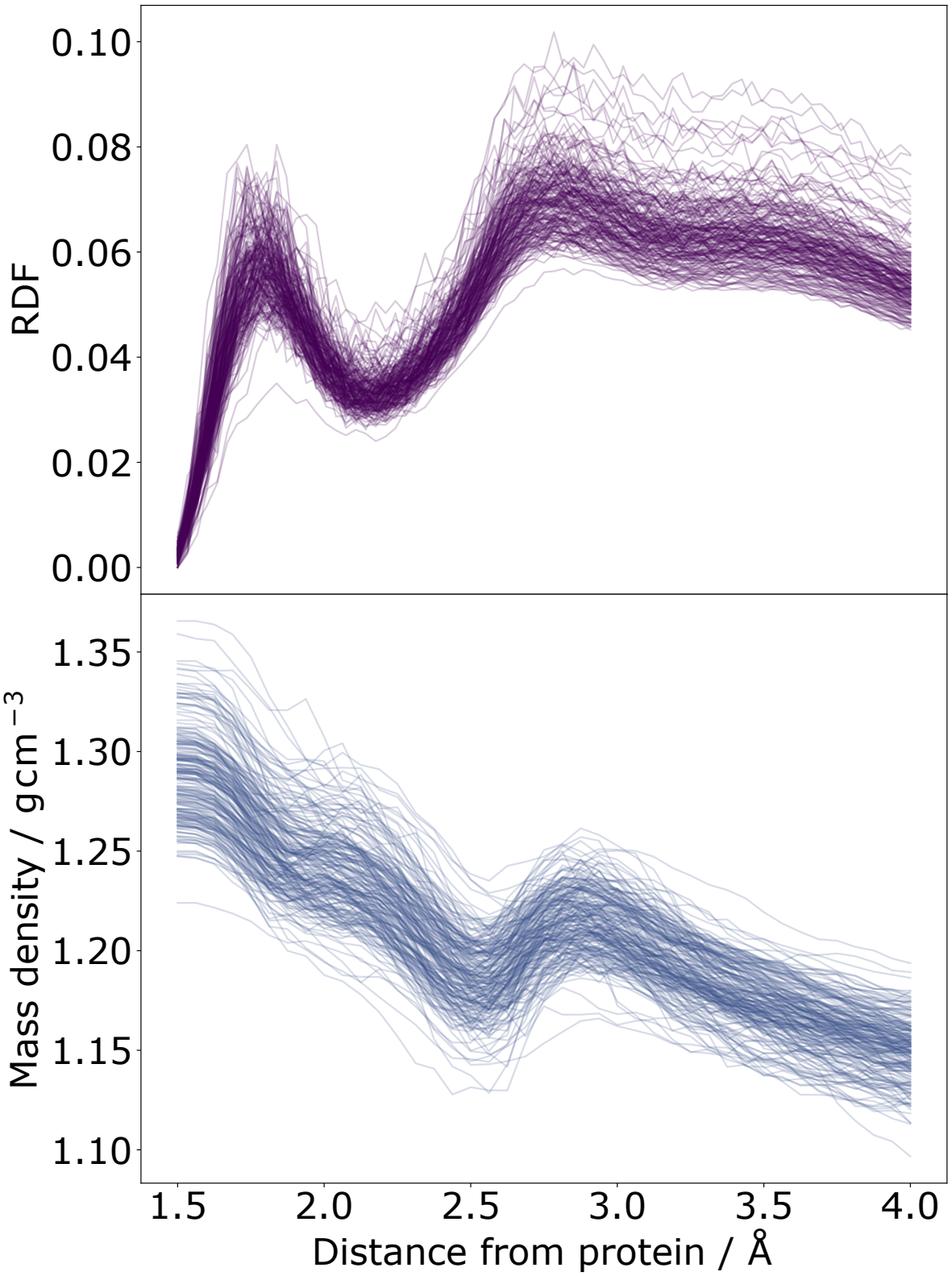

Figure S 24. **Radial Distribution Functions and protein mass density for all proteins in the protein dataset.** *The RDF and protein mass density at increasing distance from the protein, calculated by considering the mass of protein and water atoms enclosed by a volume defined by the volume of the protein plus any voxels within a varying distance from the protein. There is noise in the RDF and protein mass density signal for single structures, however, the positions of the first and second hydration shells — defined by the minima in the RDF — are consistent.*

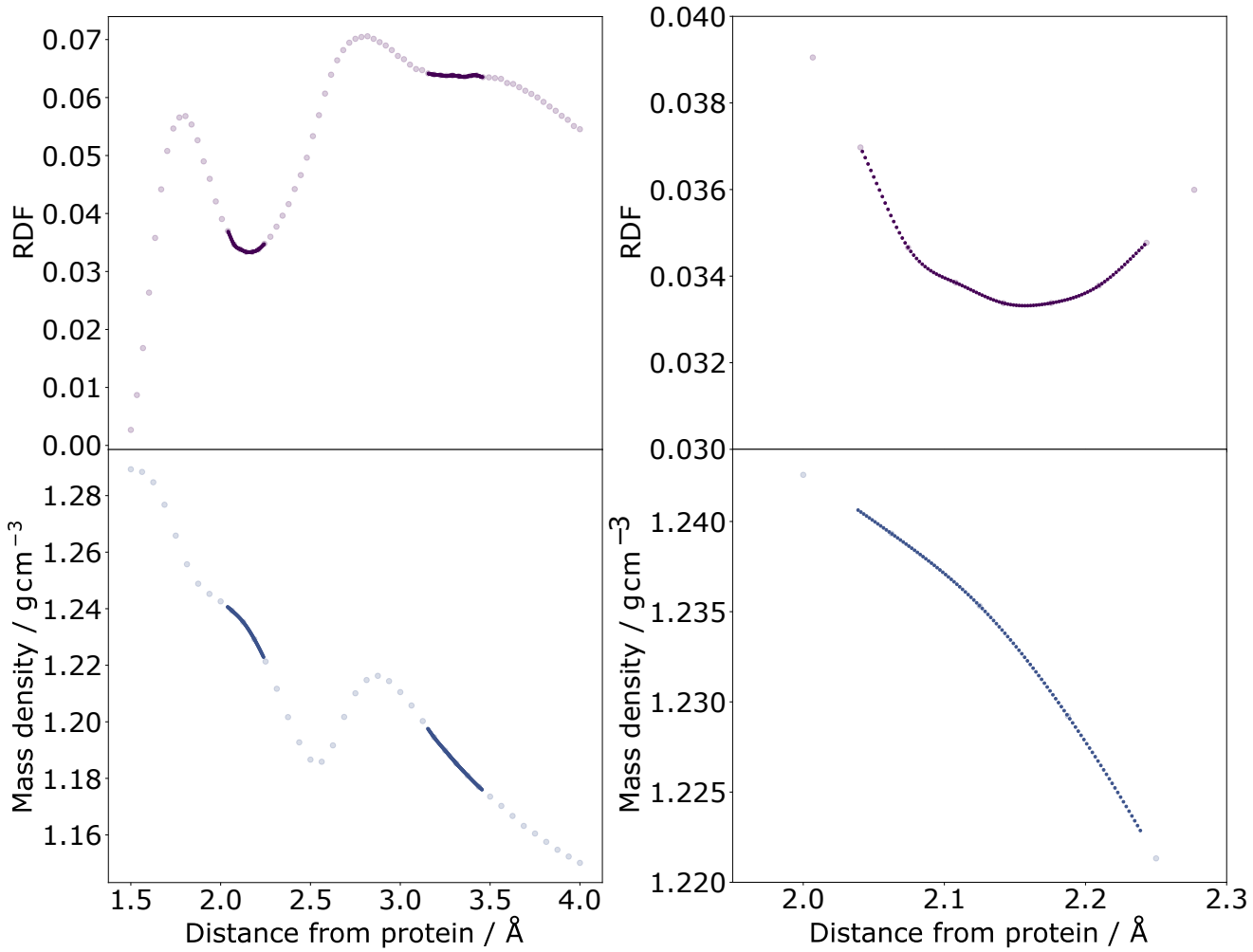

Figure S 25. **Demonstration of the interpolation around the mean position of the end of the first hydration shell.** *The light points represent the original data, whereas the opaque points represent the results of a cubic 1D interpolation of these results of 101 points within the relevant region surrounding the minimum in the RDF.*

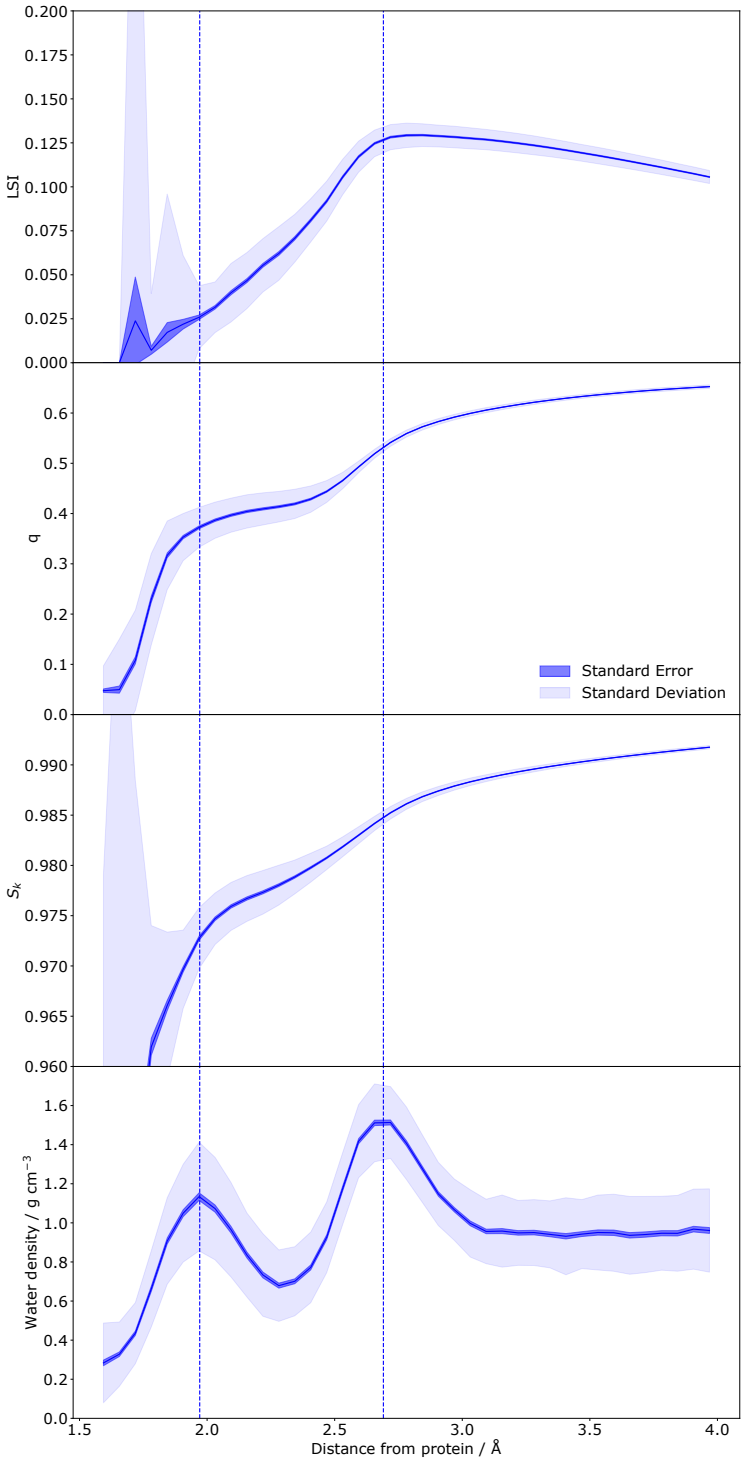

Figure S 26. **Relationship between water distance from a protein, and its order parameters and density.** *From top to bottom, we report water local structure index (LSI), orientational tetrahedral order parameter (q), translational tetrahedral order parameter (S_k), and water density (see also Figure 5 in Main Text). The centre of first and second solvation shells are indicated with vertical dashed lines at distances of 1.98 and 2.69 Å, respectively. The two shells are denser than bulk (around 1 g cm^{-3}). The second shell features peaks in LSI and q values, indicating that within this region water becomes more organized. A higher density value coupled with values of s_k lower than bulk indicates a local structure intermediate between bulk and ice.*

References

[1] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[2] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," in *Proceedings of the 15th Python in Science Conference* (Sebastian Benthall and Scott Rostrup, eds.), pp. 98 – 105, 2016.

[3] I. Alibay, J. Barnoud, O. Beckstein, R. J. Gowers, P. R. Loche, H. MacDermott-Opeskin, M. Matta, F. B. Naughton, T. Reddy, and L. Wang, "Building a community-driven ecosystem for fast, reproducible, and reusable molecular simulation analysis using mdanalysis," *Biophysical Journal*, vol. 122, p. 420a, Feb. 2023.

[4] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İlhan Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature Methods 2020 17:3*, vol. 17, pp. 261–272, 2 2020.

[5] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. D. nski, D. L. Dotson, S. Buchoux, I. M. Kenney, and O. Beckstein, "Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations," *PROC. OF THE 15th PYTHON IN SCIENCE CONF*, 2016.

[6] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with numpy," *Nature 2020 585:7825*, vol. 585, pp. 357–362, 9 2020.

[7] L. S. Rudden, S. C. Musson, J. L. Benesch, and M. T. Degiacomi, "Biobox: a toolbox for biomolecular modelling," *Bioinformatics*, vol. 38, pp. 1149–1151, 1 2022.

[8] M. C. J. Wilce, M.-I. Aguilar, and M. T. W. Hearn, "Physicochemical basis of amino acid hydrophobicity scales: Evaluation of four new scales of amino acid hydrophobicity coefficients derived from rp-hplc of peptides," *Analytical Chemistry*, vol. 67, p. 1210–1219, Apr. 1995.

[9] D. S. Kim, J. K. Kim, C. I. Won, C. M. Kim, J. Y. Park, and J. Bhak, "Sphericity of a protein via the beta-complex," *Journal of molecular graphics & modelling*, vol. 28, pp. 636–649, 4 2010.

[10] P.-L. Chau and A. Hardwick, "A new order parameter for tetrahedral configurations," *Molecular Physics*, vol. 93, no. 3, pp. 511–518, 1998.

[11] M. Kiselev, M. Poxleitner, J. Seitz-Beywl, and K. Heinzinger, "An investigation of the structure of aqueous electrolyte solutions by statistical geometry," *Zeitschrift für Naturforschung A*, vol. 48, no. 7, pp. 806–810, 1993.

[12] J. R. Errington and P. G. Debenedetti, "Relationship between structural order and the anomalies of liquid water," *Nature*, vol. 409, no. 6818, pp. 318–321, 2001.

[13] E. Duboué-Dijon and D. Laage, "Characterization of the local structure in liquid water by various order parameters," *The Journal of Physical Chemistry B*, vol. 119, no. 26, pp. 8406–8418, 2015.

[14] E. Shiratani and M. Sasai, "Growth and collapse of structural patterns in the hydrogen bond network in liquid water," *The Journal of chemical physics*, vol. 104, no. 19, pp. 7671–7680, 1996.

[15] A. Kuffel, D. Czapiewski, and J. Zielkiewicz, "Unusual structural properties of water within the hydration shell of hyperactive antifreeze protein," *The Journal of Chemical Physics*, vol. 141, no. 5, 2014.

[16] A. Bondi, "Van der waals volumes and radii," *Journal of Physical Chemistry*, vol. 68, pp. 441–451, 1964.

[17] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," *Journal of chemical theory and computation*, vol. 11, p. 3696, 8 2015.

[18] S. Izadi and A. V. Onufriev, "Accuracy limit of rigid 3-point water models," *The Journal of Chemical Physics*, vol. 145, p. 74501, 8 2016.