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

C.C.W. McAllister,[1] L.S.P. Rudden,[2] E.H.C. Bromley,[1] and M.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*

(*Electronic mail: matteo.degiacomi@ed.ac.uk)

(*Electronic mail: e.h.c.bromley@durham.ac.uk)

(Dated: 20 April 2025)

I. ABSTRACT

The density of a protein molecule is a key property within a variety of experimental techniques. We present a computational method for determining protein mass density that explicitly incorporates hydration effects. Our approach uses molecular dynamics simulations to quantify the volume of solvent excluded by a protein. Applied to a dataset of 260 soluble proteins, this yields an average density of 1.296 ± 0.001 g cm^{-3}, notably lower than the widely cited value of 1.35 g cm^{-3}. Contrary to previous suggestions, we find no correlation between protein density and molecular weight. We instead find correlations with residue composition, particularly with hydrophobic amino acid content. Using these correlations, we train a regressor capable of accurately predicting protein density from sequence-derived features alone. Examining the effect of incorporating water molecules on the measured density, we find that water molecules buried in internal cavities have a negligible effect, whereas those at the surface have a profound impact. Furthermore, by calculating the density of a titin domain and of the Bovine Pancreatic Trypsin over molecular dynamics trajectories, we show that individual proteins can occupy states with close but distinguishable densities. Finally, we analyse the density of water in the vicinity of proteins, showing that the first two hydration shells exhibit higher density than bulk water. When included in cumulative density calculations, these hydration layers contribute to a net increase in local solvent density. Overall, we find that proteins are less dense than previously reported, which is offset by their ability to induce a higher density of water in their vicinity.

PACS numbers: 87.15.kr

II. INTRODUCTION

Protein mass density — the density of an individual protein molecule — is an important fundamental biophysical quantity relevant to a wide range of experimental techniques, including X-ray crystallography and ultracentrifugation studies of protein oligomers[1]. Additionally, the precise determination of protein and solvent densities is relevant to methods for which the local dielectric permittivity is experimentally implicated, including Extraordinary Acoustic Raman Spectroscopy (EARS) [2]. The protein mass density has been approximated since at least the late 1960s [3] to be effectively a constant independent of the protein's size, shape, or other physical characteristics, a consequence of the closely packed interiors of proteins [4]. Specifically, a value of 1.35 g cm^{-3} has been commonly used [5] based on early compressibility [6] and sedimentation [7] studies, though various values deriving from experimental and theoretical approaches have been proposed.

Key to the calculation of protein density is the calculation of the protein volume. There are many different definitions of protein volume, including: the geometric volume, which is the solvent-excluded volume enclosed within the solvent-excluded surface [8], also known as the molecular surface volume [9]; the van der Waals volume, the volume of overlapping spheres representing the van der Waals radii of each constituent atom of the protein [10]; and the solvent-accessible volume, the volume enclosed by the solvent-accessible surface area (SASA) [11]. This work is concerned with the molecular surface volume. This is the relevant measure of volume occupied by a protein for experiments which depend on changes in the local permittivity of proteins in solvent (e.g., EARS).

While protein mass density is often taken as a constant, in 2004 Fischer *et al.* [12] examined previously published experimental [6,7] and theoretical [5,13,14] estimates of protein densities and concluded that for relatively small proteins (below 20 kDa) protein density is molecular weight dependent, in an inverse exponential relationship. The authors further proposed that, for larger proteins, a constant value of 1.410(6) g cm^{-3} should be considered. Theoretical calculations of protein density used in the meta-analysis of Fischer *et al.* were performed on crystal structures almost completely devoid of water molecules using Voronoi Tessellation methods (Andersson and Hovmöller, 1998 [5]; Tsai *et al.*, 1999 [14]; Quillin and Matthews, 2000 [13]). These techniques provide analogous approximations of a dry molecular surface volume. Crucially though, water molecules form hydration shells around proteins, whereby coordinated waters essentially behave as an integral part of the protein [15]. The specific distribution of conformational states of protein side chains depends on their complex interactions with these water molecules, thus the evalua-

tion of protein-water dynamics is required to accurately calculate protein volume. Water-protein interactions play key structural roles in proteins, driving their organisation and flexibility, and ultimately impacting upon their function [16,17]. For this reason, computational methods that can accurately incorporate hydration into the calculation of protein density would be advantageous. Multiple methods to calculate the molecular surface volume exist including analytic methods like MSMS (Michel Sanner's Molecular Surface), which can compute the molecular surface volume via a procedure that relies on the reduced surface [18], and explicit surface representation methods like LSMS (Level Set method for Molecular Surface generation), which uses a level-set front-propagation method [19]. Inferring protein volumes from their amino acid sequence, with the volume calculated as the sum of the volumes of the constituent amino acids, has also been shown to produce remarkably accurate results [20]. A recent study of the partial specific volumes of proteins, the inverse of the protein density in solution, using this method has calculated a theoretical mean value for all human proteins of 0.735 mlg^{-1} with a standard deviation of 0.010 mlg^{-1}, equivalent to 1.36 $\pm$ 0.03 g cm^{-3}, and an approximately Gaussian distribution [21]. However, none of these methods account for the effects of the protein's hydration shell. Furthermore, the partial specific volume is a macroscopic experimentally observable quantity from which the mass density of individual protein molecules cannot necessarily be simply derived.

Previous computational studies of protein mass density that have attempted to account for the surface effects of water have mostly utilised solvent-corrected Voronoi tessellation methods [5,13,14]. While useful, the Voronoi method is known to produce inaccurate results for surface atoms which are sparsely surrounded by other atoms, leading to different density results depending on how or if the contributions of these surface atoms are accounted for [13]. The Voronoi method as applied to proteins is usually also adjusted to account for different atomic radii which can introduce vertex errors, though these can be eliminated by enveloping each Voronoi cell with a hyperbolic surface [22]. While some previous work has used Voronoi methods including water molecules from a simulation, dividing the entire simulation box including solvent into Voronoi polyhedra [23], most have focused on calculating volumes from crystal structures.

In this work we consider protein density to be associated with the volume of solvent they displace, whereby the position of water molecules coordinated with the protein is explicitly determined and accounted for using all-atom molecular dynamics (MD) simulations. To this end, we adopt a voxel-based method analysing the position of protein and water atoms in molecular dynamics simulation snapshots. This approach offers key advantages. Firstly, utilising MD simulations allows for volumes and densities to be calculated for many protein conformers, yielding better statistics than when single atomic structures are used. Second, it avoids assumptions that may affect density estimation, e.g., water has homogeneous density everywhere and is ideally packed around every protein atom. Third, it enables characterising how changes in environmental conditions (e.g., temperature, pressure, pH, solvent composition) might affect protein density. Hereafter we present our method, before using it to determine protein density in solution for a large set of proteins and assessing whether this property depends on any physical characteristics. Furthermore, to test the assumption of close internal packing leading to a constant protein mass density value [4], we present a method to identify the presence of buried water molecules within the protein interior.

We use our method to calculate the density of a dataset of 260 soluble proteins. We then train a random forest regressor (RFR) [24] with our calculated densities and a range of structural features, demonstrating that soluble protein density can be predicted from amino acid sequence alone. Finally, we examine how mass density might vary within individual proteins in both equilibrium and non-equilibrium conditions. To this end, we examine two case studies: Bovine Pancreatic Trypsin Inhibitor (BPTI) and immunoglobin-like (Ig-like) domain of titin. BPTI, the first protein simulated with molecular dynamics [25], has a well-characterised conformational space thanks to a 1-millisecond unbiased simulation performed by DE Shaw et al. [26]. Here, we utilise Markov State Modelling (MSM) to divide this simulation into discrete states and show that these feature distinct densities. Titin contributes to the passive elasticity of muscle by acting as a molecular spring [27]. It is the largest known protein consisting of up to 300 mostly Ig-like domains [28] which unfold sequentially under the influence of an external stretching force, with the domains refolding upon relaxation [29]. Utilising a steered molecular dynamics (SMD) simulation of a single titin Ig-like domain, we evaluate the evolution of protein mass density over the large-scale conformational change induced by mechanical stress. We find that within the SMD simulation there exists two sub-populations — divided by secondary structure content — with different densities.

Finally, we extend our density calculation method to the hydration shell surrounding a protein, and use it to investigate the structure of water around each protein in our dataset. In agreement with previous experimental [30] and simulation [31] data, we find that the mean first hydration shell density (1.1 $\pm$ 0.3 g cm^{-3}) is 12% greater than bulk water. Additionally, we find that the second hydration shell is on average significantly more dense (1.5 $\pm$ 0.2 g cm^{-3}, with a density increase of 54.5% compared to bulk water. The unexpected extent of this increase in the presence of a protein molecule may explain the discrepancy between experimental and computational estimates of protein density.

III. METHODS

A. Protein dataset

To study the density of soluble proteins, we took a diverse set of 260 protein monomers featured in the protein-protein docking benchmark 5 [32]. To accurately measure their density, linked to the volume of water they displace, we aim to explicitly determine how water molecules arrange themselves around each protein atom. To this end, we solvated each pro-

tein in a TIP3P water box neutralized with Na^+ and Cl^- counterions, and relaxed the resulting system using molecular dynamics (MD) simulations using the GROMACS engine and the Amber ff14SB [33] force field. Proteins were first energy minimized using a steepest descent algorithm until a maximum force of less than 1000 kJ mol^{-1} nm^{-1} was achieved. Then, they were all equilibrated in the NVT ensemble at a temperature of 300 K, using a 2 fs timestep with bonds restrained using LINCS. A particle-mesh Ewald [34] summation was used to treat long-range interactions and the velocity-rescaled modified Berendsen temperature coupling method [35] applied separately to protein and non-protein atoms. Finally, 1 ns production runs were carried out in the NPT ensemble, with 300 K and 1 bar determined by modified Berendsen temperature coupling and Parrinello-Rahman pressure coupling [36]. Protein-water conformations were extracted every 50 ps from each production run, leading to the extraction of 5460 snapshots (21 for each protein). Since water density is itself temperature-dependent, to assess the effect of temperature on measured density, we repeated the simulation protocol described above, for all proteins, at a physiological temperature of 310.15 K. Finally, to ensure consistency across force fields and water models, we also repeated our simulation protocol at both 300 and 310.15 K using the Amber ff99-ILDN [37] force field and the SPC/E water model (see Supplemental Materials). The specific force field and water model combinations used in this work were chosen for the reported agreement between protein hydration shell contrasts predicted using them, and experimental small angle X-ray and neutron scattering data [38]. Unless otherwise specified, in the main text we report results obtained at 300 K using the Amber ff14SB force field and the TIP3P water model.

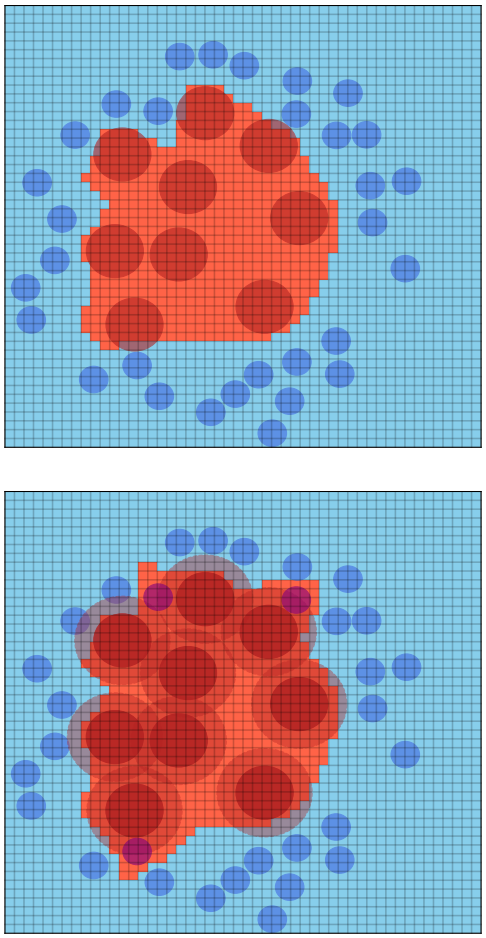

FIG. 1. Schematic representation of our protein volume calculation method. Red and blue circles represent the van der Waals radius of protein and water atoms, respectively. The space surrounding the protein is divided in a fine grid, coloured according to which atom is the closest. (top) the red region represents protein occupied volume. (bottom) the red region represents the volume occupied by protein and solvation shell. This is calculated by considering as part of the protein any water atom within a given cutoff distance from the centres of any protein atoms (transparent circles surrounding red circles). This region may be incrementally expanded to allow for the characterisation of changes in water density at different distances from the protein.

B. Protein volume calculation

To calculate the volume occupied by a protein, we place an equilibrated MD simulation snapshot of the protein with its surrounding water into a 3-dimensional grid (Figure 1) defined by two parameters: the size of its cubic voxels (hereafter "step"), and the amount of extra space added to the grid at the extremities of the protein in each Cartesian axis ("leeway"). Our method is implemented in Python, using NumPy [39], SciPy [40], and MDAnalysis [41,42] packages (see Supplementary Information).

For each layer of the grid corresponding to a specific Z-value in Cartesian space we calculate the distances from the centre of each voxel to all protein and solvent atom positions. The distance from the centre of each voxel in the layer to the nearest protein and solvent atoms is then calculated, and every voxel for which the nearest protein atom is closer than the nearest solvent atom is considered part of the protein, with a voxel of volume (simply equivalent to the step size cubed) added to the total calculated volume. By performing the calculation layer-by-layer the memory requirements of the procedure is drastically reduced (see Figures 1 and S1). To make computation more efficient a 'shell' parameter was added to define a shell around the protein (or a distance from each protein atom) outside of which water molecules would be excluded from the calculation. We found that a shell distance of 6 Å increases computational efficiency without sacrificing accuracy.

To calculate the density of a protein combined with adjacent water molecules within a cutoff distance, we adapted our algorithm to enable considering selected water molecules as part of the protein itself. To assess how water is distributed around a protein, we calculate the radial distribution function (RDF) between all protein atoms, and all solvent ones. The locations of the first and second minima in the RDF are identified to give the thickness of the first and second hydration shells [43].

Our density calculation algorithm can accommodate the usage of atoms' van der Waals radii, by subtracting the atomic radii from the calculated distances prior to evaluating whether a grid point is closer to protein or water atoms. We assessed the effect of this improvement by either using radii from *A. Bondi, 1964* [44], or from the Amber ff14SB force field. Overall, we found that accounting for van der Waals radii has a minimal effect on the calculated volumes. Therefore, to reduce computational time, we consider all atoms as having the same radius (see Figures S6 and S7).

The main parameters for the protein volume calculation are the step size of the grid, the leeway of the grid, and the tolerance distance for water molecules to be considered in the calculation (the 'shell' distance). The accuracy of our algorithm can also be increased by averaging density measurements over a number of rotations of the protein-water system in relation to the grid. We evaluated the accuracy and computational cost of different combinations of parameters (leeway, grid size, shell distance, number of rotations, see SI and Figures S3, S4, and S5). We found that rotating the protein-water system was less computationally efficient than reducing the step size. Thus, we used a step size of 0.5 Å, shell distance of 6 Å, and a leeway of 5 Å, without any system rotation (see SI for details).

C. Residue volume calculation

Similarly, it is possible to calculate the volume of an individual residue of a protein by determining which voxels are closer to the atoms of this residue than to the atoms of any other part of the system (i.e., the atoms of all other residues of the protein and the atoms of the solvent). The sum of these residue volumes for a single protein should be similar to the protein's volume, as calculated previously, with slight discrepancies due to the non-global nature of this type of calculation possible. To reduce such discrepancies, the voxel grid was defined in the same way for the global protein volume calculation and for the residue volume calculation, with the relevant sections of the surrounding voxel grid assigned to each residue in the latter case. In practice, a small discrepancy of 0.024% is observed (see SI for details).

D. Excess protein volume

Characterising the volume of a protein by the volume which it displaces in solvent, specifically defined at the surface of the protein by the midway point between protein atoms and the nearest solvent atoms, necessarily results in a region of what we define as protein volume being outside of the van der Waals radii of either protein or solvent atoms. This region, surrounding the van der Waals surface of the protein, can be estimated by calculating the volume of voxels which are both outside the van der Waals radii of their nearest protein or solvent atoms and at the solvent-interface surface of the protein. This excess volume can then be optionally removed to give a protein volume that matches the van der Waals protein volume at the protein's surface.

E. Identification of buried waters

To identify waters that are buried inside the protein (which would typically not be considered part of the protein, but are likely to be an integral part of it) [45], we used the DBSCAN clustering algorithm [46] as implemented in the scikit-learn Python library [47] to cluster water molecules by the coordinates of their oxygen atoms. Using a minimum sample size of 1 and 4.0 Å as the maximum distance allowed between samples within the same cluster, the first cluster (sorting by cluster size) represents the bulk water, including the hydration shells. Therefore, further clusters of water molecules tend to describe water molecules buried inside the protein, either alone or in groups. By filtering water molecules identified as being part of these clusters by ensuring that they are located within 3 Å of a protein atom reliably identifies water molecules that are buried within the protein. Optionally, our method enables calculating the protein density by considering internal water molecules as part of the protein.

F. Analysis of water density

The density of the protein-solvent system can be calculated while successively including more solvent moving out radially from each protein atom, with all solvent atoms within a specific distance from the centres of each protein atom becoming part of a protein-solvent complex and thus considered in the density calculations (see Figure 1). By excluding the protein's mass and volume contributions to this calculation, it is possible to calculate water density as a function of distance from the protein-water interface.

G. Protein physical properties determination

We extracted various protein physical characteristics to evaluate any correlation between them and either the protein mass density, or the change in protein mass density upon including the effects of buried waters. These characteristic values include SASA, sphericity, aspect ratio [48], amino acid composition, and protein molecular weight. The SASA is calculated via the Shrake-Rupley algorithm [49], otherwise known as the "rolling ball" method. A probe size of 1.4 Å is usually used to represent a water molecule, an approximation of the water molecule's van der Waals radius (more accurately, half of the oxygen-oxygen distance between two hydrogen-bonded water molecules [50]). The amino acid compositions of surface and interior (i.e., non-surface) parts of the proteins were also calculated, with protein atoms being defined as surface atoms if more than 5% of the spherical mesh points surrounding each atom were found via the SASA algorithm to be surface accessible to a 1.4 Å probe.

We assessed the normality of distribution of protein density and each physical characteristic for normality using the Shapiro-Wilk Test for normality [51]. We calculated Pearson's correlation coefficients between the protein mass density (which was found to be approximately normally distributed)

and physical characteristics. As not all physical characteristics were found to be normally distributed, we also evaluated the Spearman correlation coefficient, which is more appropriate in these circumstances (see Tables S1 and S2).

H. Random Forest Regressor

To predict protein densities, we trained random forest regressors (RFR) using the scikit-learn Python library [47]. We built two random forest regressors: one trained using only sequence-derived features (seq-RFR) and another trained using sequence and structural features (struct-RFR). The features included in the latter included amino acid residue composition, secondary structure (helix, strand, and coil percentages), net charge, and the prevalence of hydrophobic residues. We optimised the regressors' hyperparameters (maximum depth, minimum number of samples per leaf, and number of estimators) via a grid search with cross-validation, measuring the performances of cross-validated models by their R^2 score to ensure the residuals are minimized. For struct-RFR, we also carried out an ablation study to identify the minimal set of features leading to highest performance in the classifier, which yielded a classifier operating on 20 features. For details on selected features, and training and validation protocols, see Supplementary Methods, Tables S1 and S2, and Figures S17-22.

I. Markov State Modelling of BPTI

A down-sampled version of the 1-millisecond BPTI simulation performed by DE Shaw et al. [26], with a timestep of 10 ns, was used to determine distinct conformational states of the protein via a Markov State Modelling (MSM) analysis using the PyEMMA Python package [52].

Dimensionality reduction was performed using Time-lagged Independent Component Analysis (TICA) [53,54] of the α-carbon Cartesian coordinates. The TICA coordinates were subsequently clustered using κ-means clustering [55] to produce 300 discrete clusters. We verified the estimated Markov State Model via an well-established procedure [56]. First, by analysing its implied timescales, and by performing a Chapman-Kolmogorov test (see Figure S23). Perron-Cluster Cluster Analysis ++ (PCCA++) [57,58] algorithm is then used to assign a probability for each κ-means state being a member of a smaller collection of 5 metastable macrostates. We then find the κ-means cluster with the highest probability of being in each metastable PCCA++ state and sample 50 trajectory frames that are associated with each, for a total of 250 protein conformations over 5 metastable states. All atom protein conformations were then solvated with TIP3P water, simulated with the Amber ff14SB force field, and had their density calculated via the routine previously detailed for the 260-protein dataset.

J. Steered molecular dynamics of titin

The I27 domain of titin (PDB: 1TIT) was aligned so that the vector connecting N- and C-terminus lays on the x-axis, and solvated in TIP3P water (155x60x62 Å). The resulting box was neutralised with Na^+ counterions, and simulated with the Amber ff14SB force field, using a 2 fs time step, and PME handling long range electrostatic interactions. The system was first energy minimised for 500 steps, then simulated for 50 ps in the NVT ensemble, with 300 K imposed via Langevin dynamics (damping of 5 ps^{-1}) and α-carbons restrained with a harmonic potential of 10 kcal mol^{-1}. Maintaining the restraints, 100 ps were then simulated in the NPT ensemble with 1 Atm imposed by a Langevin piston (period of 200 fs, decay of 50 fs), followed by 1 ns without restraints. From this equilibrated state, the unfolding of titin subject to mechanical stress was simulated via steered molecular dynamics (SMD). To this end, the N-terminus of titin was restrained with a harmonic potential, while the C-terminus was pulled for 8 ns at a constant velocity of 25 Å ns^{-1} along the vector connecting the termini (thus extending within the elongated water box), with a force constant of 7 kcal mol^{-1} $Å^{-2}$.

From the SMD, we extracted titin conformations every 20 ps between 1 and 7 ns, for a total of 291 titin conformations. To ensure water is correctly packed around each extracted conformation, removing effects that might be caused by a fast steering or imposed restraints, each system was re-solvated and relaxed for 1 ns without any restraint, utilising the procedure previously outlined for the 260-protein dataset. For each of these individual simulations, We extracted snapshots at 0, 0.5 ns and 1 ns, for a total of 873 solvated titin conformations. For each of those conformations, we calculated density and percentage of amino acids part of a β strand calculated in MDTraj [59] using an implementation based on DSSP-2.2.0 [60]. We used the statistical distribution of this latter quantity, bimodal in nature, to subdivide the ensemble of titin conformers in two sub-populations using the minimum between its two peaks as classification criterion.

IV. RESULTS

A. Mean protein density

For each of the 4221 protein-solvent conformation extracted from each of 260 simulations of proteins in water, we identified internal water molecules for 88.2% of structures (85.5% when averaging over multiple simulation frames), with an average of 14 ± 8 buried water molecules identified. However, including these buried waters in our calculations had a statistically insignificant effect on the calculated protein density (see Figure S8). This is explained by the small volume occupied by a single water molecule (30 $Å^3$) [61], and by the number of internal waters varying linearly in proportion to the molecular weight At 300 K, we calculated a mean protein density of 1.294 $\pm$ 0.004 g cm^{-3} when including the effects of internal water molecules, and 1.296 $\pm$ 0.001 g cm^{-3} when neglecting them. Our benchmarks also showed that changing the simula-

tion temperature with the Amber14SB force field and TIP3P water model combination had only a minimal effect on the calculated densities (mean densities of 1.295 ± 0.002 g cm^{-3} at 300 K and 1.293 ± 0.002 g cm^{-3} at 310.15 K). Conversely, altering the simulation temperature yielded significant effects when using the Amber 99SB-ILDN force field and SPC/E water model combination, whereby at 310.15 K the mean protein density decreased to 1.289 ± 0.002 g cm^{-3} from 1.298 ± 0.0004 g cm^{-3} at 300 K (see Figures S9 and S10).

B. Correlation of density with other physical properties

While we did not find any correlation between protein mass density and protein mass (PCC: 0.02446, p-value: 0.7265, SCC: −0.04361, p-value 0.5326), many physical characteristic values of proteins did have small but significant correlations with the protein mass density (see Figures S11-S15). The overall protein charge, the percentage of charged residues, and the percentage of hydrophobic residues were found to have weak correlations with the protein mass density, with slightly larger correlations resulting from considering the amino acid composition of the surface and interior of the protein separately (SCCs ranging from −0.1672 to −0.5142).

Aggregating the mass densities of all the protein-solvent conformations in our dataset we found a mean value of 1.289 ± 0.002 g cm^{-3}. The inclusion of buried water molecules in the density calculation reduced the calculated density for all protein-solvent conformations for which buried waters were identified, though the overall effect was so small that the sample mean remained unchanged. For proteins with multiple conformers, the mean standard deviation of the protein mass density was 0.006 g cm^{-3}, suggesting that dynamics in the sub-ns timescale had a limited effect on the protein mass density. While fast dynamics have limited effect on protein volume (and hence protein mass density), slower dynamics over longer timescales might still have a significant effect.

C. Protein density prediction

We investigated whether the density of a protein can be predicted based on a collection of structural and sequence features. To this end, for each protein we measured 40 structural features, and amino acid composition (20 features quantifying the percentage presence of each amino acid in the protein sequence). We then trained Random Forests regressors (RFR) with a combination of the structural and sequence features (struct-RFR), or with sequence features alone (seq-RFR). In both circumstances, we found that the regressors were able to accurately predict protein densities, with seq-RFR (mean squared error (MSE): 3.88e-0.5, Pearson correlation coefficient (PCC): 0.967) outperforming struct-RFR (MSE: 2.19e-05, PCC = 0.981). This means that protein densities can be quickly and accurately predicted from only a protein's amino acid sequence. Reviewing the importance of each feature, we found that protein mass is one of the least important features (see Figure S21). The most important features for struct-RFR are the percentage of aliphatic hydrophobic residues, the percentage of hydrophobic residues, and the total protein charge.

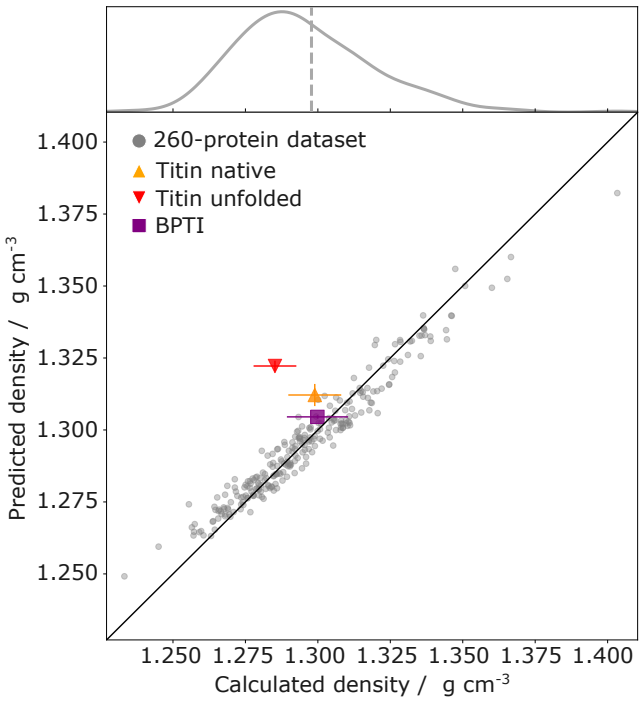

FIG. 2. Comparison between calculated densities, and densities predicted by a Random Forest Regressor (struct-RFR). The RFR was trained on a set of 20 sequence- and structure-based features gathered from our 260-protein dataset, and the equilibrated crystal structure of BPTI (PDB: 5PTI) and titin (PDB: 1TIT) in its folded and extended state. The identity line is shown in black for comparison. We find a PCC of 0.976 for the 260-protein dataset. The RFR accurately predicts the densities of conformations of BPTI (of which one structure is present in the training set) and folded titin (not part of the training set). The density of mechanically unfolded titin conformations are poorly predicted. Above, a Kernel Density Estimation plot of the distribution of mean calculated densities in the 260-protein dataset, with the overall mean density of 1.296 ± 0.001 g cm^{-3} annotated with a dashed vertical line.

D. Density variation in individual proteins

To evaluate the extent of density variations within individual proteins, we studied two different cases, a long equilibrium simulation of BPTI, and a non-equilibrium simulation reproducing the unfolding of an Ig-like domain of titin under mechanical stress. BPTI is featured in the training set, whereas the titin domain is not (see Figure S2).

For BPTI, we found that the simulation is divisible into five metastable states with distinct densities and interior aliphatic hydrophobic residue prevalences (see Figure 3). These are associated with structural differences in the N-terminal 3_{10} helix — which is less prevalent in the conformers of states 1 and 2 — and different arrangements of the loops between residues 7-16 and 35-46, leading to different degrees of compaction.

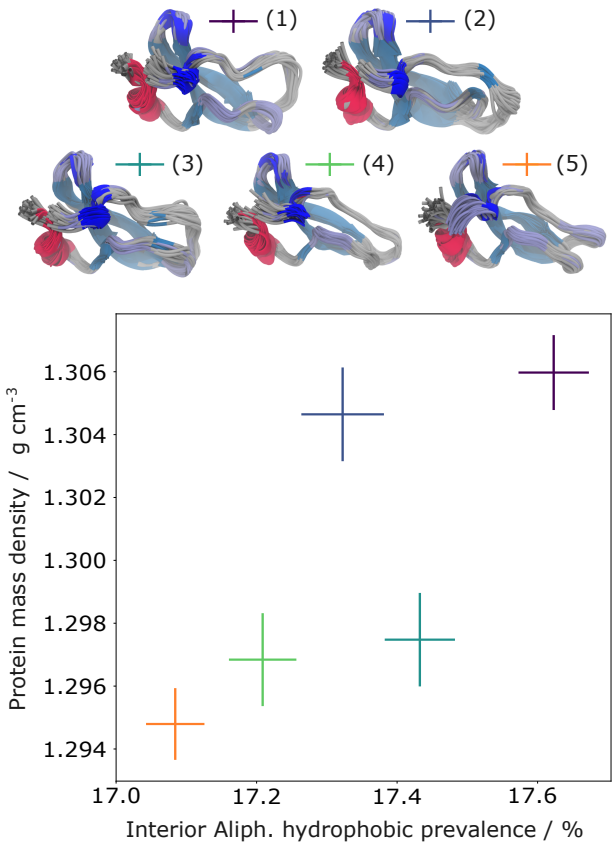

FIG. 3. BPTI features states with distinct densities. A 1 ms molecular dynamics simulation of BPTI can be subdivided in five metastable states via Markov State Modelling (each represented by 50 overlaid conformers at the top). The secondary structure is coloured as: α-helix in red, 3_{10}-helix in blue, β-sheet in turquoise, turns in light blue and coil in grey. These conformers differ in their degree of compaction, as captured by the prevalence of aliphatic hydrophobic residues in their interior and their density. The structures of states 4 and 5 most closely resemble the crystal structure of BPTI (PDB: 5PTI), with a difference of only a more significant turn in the coil region of residues 12 and 13 in state 5. States 1 and 2 are differentiated from the others by the loss of structure of the small N-terminal 3_{10}-helix, though for some conformers this feature is intact.

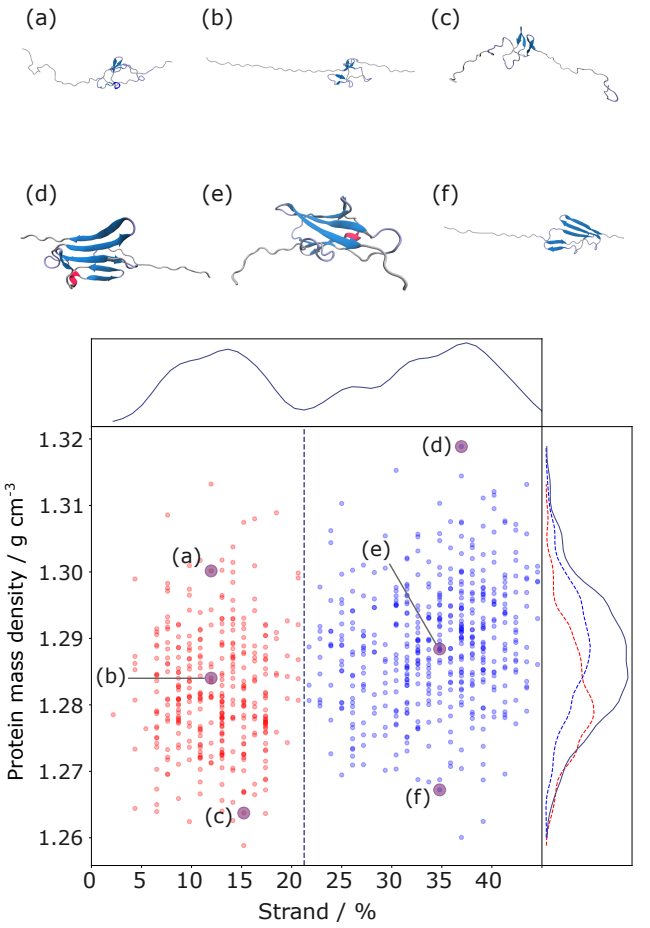

FIG. 4. Unfolding of titin under mechanical stress reveals conformer sub-populations with distinct densities. For each titin conformation in a steered molecular dynamics simulation (see example conformers at the top), we calculate density and percentage of preserved secondary structure. The distribution of secondary structure content (kernel density estimation in the upper graph) reveals two distinct sub-populations, identified with red and blue colours in the scatter plot. In the right graph, kernel density estimations reveal that these sub-populations feature distinct density distributions (red and blue lines). The density distribution of the whole simulation is shown in black.

For titin, we found that the 873 equilibrated conformations extracted from the SMD simulation could be separated into two distinct sub-populations based on the percentage of residues found in a β-strand conformation. The sub-population associated with lower strand percentage (i.e., the more unfolded conformations, red in Figure 4) has a mean density of 1.2819 $\pm$ 0.0005 g cm^{-3}. The sub-population associated with higher strand percentage (i.e., conformations closer to the native state, blue in Figure 4), has instead a marginally higher, though distinct (t-test statistic: -11.7, p-value: 1.61e-29), mean density of 1.2892 $\pm$ 0.0004 g cm^{-3}.

Importantly, while both BPTI and titin feature conformations of varying density, we found that these variations fall within the distribution of densities of proteins with comparable physical properties in our protein training set (see Figures 2 and S2). We challenged our trained struct-RFR with snapshots from the simulations of BPTI and titin, subdivided in its folded (native) and unfolded (extended) subpopulations. Density predictions for BPTI conformers were accurate overall, titin native conformers were slightly overestimated, whereas extended titin conformations were incorrect (see Figure 2). This failure is not unexpected, given that the training set consisted only of proteins in their folded native state. Qualitatively, this failure is explained by the fact that the most important feature for struct-RFR is the prevalence of aliphatic hydrophobic residues interior, a quantity which is negatively correlated with protein density. Hence, when titin unfolds and the interior aliphatic hydrophobic residues become exposed to solvent, the internal prevalence of these residues decreases, leading the RFR to predict a lower density than we calculate.

E. Effect of hydration on protein density

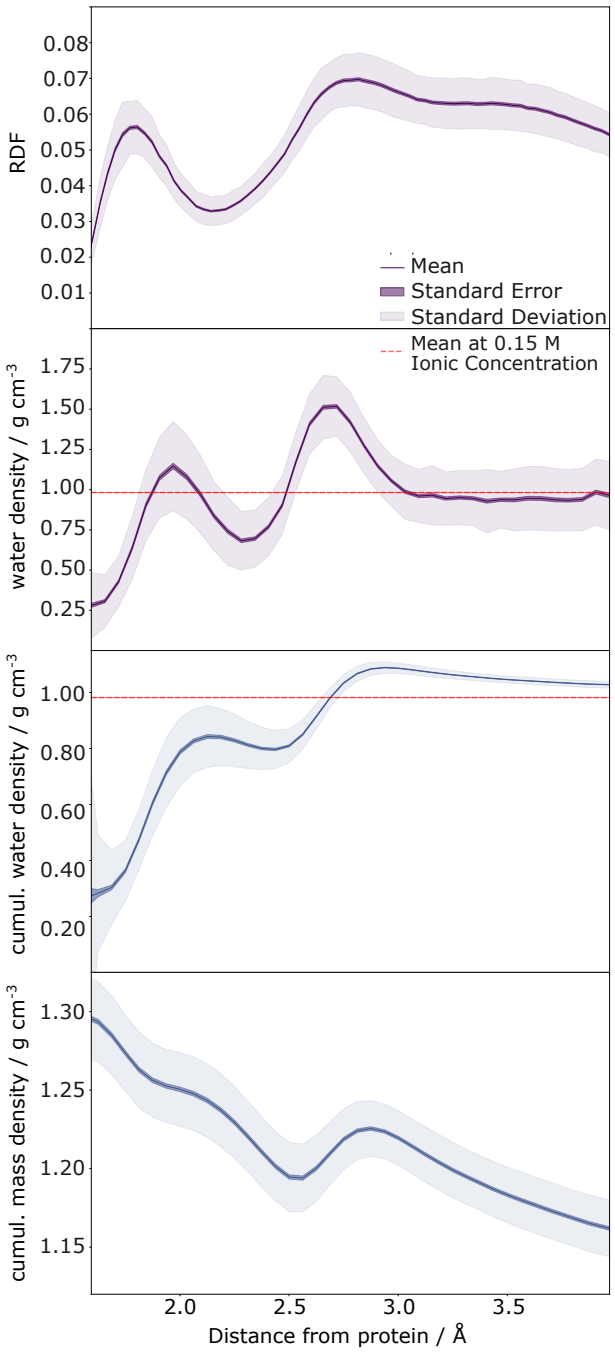

FIG. 5. Relationship between protein and water density, averaged over the whole protein dataset. The two top graphs, in palatinate colour, report on water radial distribution function (RDF) and water mass density. The bottom two graphs, in blue, report on the cumulated effect on measured mass density of water, or the combined protein-water system, when accounting for an increasingly large water shell around the protein. The effective protein-water mass density decreases the more water is included, with a non-monotonical trend determined by water having density higher than bulk value in the first two hydration shells.

Water forms hydration shells around solutes, with water molecules in contact with the protein featuring dynamics more akin to those of the protein, than those of bulk water [62]. We therefore investigated how the measured protein density might be altered in situations whereby water molecules in the immediate vicinity of the protein are included in the calculation (see Figure 5 and S24). Averaging over all proteins in our dataset, we obtained a mean thickness of the first hydration shell of 2.197 $\pm$ 0.001 Å, with a standard deviation of 0.01 Å. Including the first hydration shell in our calculations led to a mean protein mass density of 1.228 g cm^{-3} , while including the second hydration shell (found at a cutoff distance of 3.322 $\pm$ 0.004 Å) led to a density of 1.185 g cm^{-3}. Interestingly, we observe that the density of a protein-water system decreases non-monotonically when an increasing amount of water surrounding the protein is included. This indicates that the density of water surrounding the protein is not constant.

F. Hydration shell structure

Our results show that the presence of a protein molecule can significantly alter the density of the water that surrounds it. So, we finally quantified how the presence of a protein might affect adjacent water structure (see Figure 5). We found that, for the first and second solvation shells, water is denser than bulk. Specifically, on average water reaches a density of 1.1 $\pm$ 0.3 g cm^{-3} in the first shell (12% greater than bulk water), and an even larger density of 1.5 $\pm$ 0.2 g cm^{-3} in the second shell (54.5% denser than bulk). Investigating the order of water at a range of distances from proteins (see Figure S26) we also observed that this becomes more organized with successive solvation shells. Considering the density of an increasingly large shell of water around a protein, we observe that if at least two water shells are present, the average water density will be greater than bulk, slowly converging to bulk if more water is accounted for.

V. DISCUSSION AND CONCLUSION

In this work we have produced a large dataset of hydrated protein structures via molecular dynamics simulations to accurately assess the mass density of proteins, establish any possible correlations between mass density and physical characteristic values of proteins, determine the effect of the inclusion of hydration shells on the apparent protein mass density, and investigated the effect of the protein on the organisation of the water around it. To determine protein mass density, we developed and profiled an efficient voxel-based method, which is also able to identify and account for buried waters.

For our dataset, we calculated a protein mass density of 1.296 $\pm$ 0.001 g cm^{-3} at 300 K. These measures are essentially unaffected by the presence of buried water molecules, and only marginally increased when using a different force field and water model. Overall, the values we measured are lower than the 1.35 g cm^{-3} value commonly used in the scientific literature. Furthermore we found that, in contrast to previous

research, there is no correlation between protein density and mass. However, we identified other physical characteristics that are significantly correlated with the protein mass density. These include the overall charge, the percentage of hydrophobic amino acid residues, and the percentage of charged surface amino acids. We also demonstrated that these correlations can be exploited by a Random Forests regressor to predict with high accuracy the densities produced by our MD-based method, at a fraction of the computational cost. Remarkably, we also found that a regressor could yield high quality predictions based on the amino acid sequence alone.

As proteins are dynamic in nature, we also investigated how the density of an individual protein might evolve using molecular dynamics simulations of BPTI and titin. While our regression model demonstrated that the main determinant of density in a protein is amino acid composition, our results show that conformational changes also have a measurable effect. Given that such dynamics-dependent effects are subtle, we expect an amino acid composition-dependent density value to be a suitable proxy for most experiments. However, the variations we observed highlight the presence of volume changes at biologically relevant frequencies, which might be identifiable with techniques sensitive to fluctuations in the distribution of charges and electron density in an analyte.

Finally, we characterized how the hydration shells affect the resulting measured density of both protein and water. The non-monotonic decrease in the density of the protein–water system when increasing the number of water molecules included is due to the varying density of water in the first two hydration shells. Indeed, we found that the hydration shells have a density higher than the bulk, with the second shell exceeding the density of the first. We found that including the first hydration shell in the protein-water nanobioparticle resulted in a mean protein mass density of 1.252 g cm^{-3}, while including the second hydration shell further reduced the measured mass density to 1.192 g cm^{-3}. These values are relevant for any experiment studying the properties of proteins in solution. Overall, this observation highlights how considering water as bulk around an analyte of interest might affect the quantities extracted from a measurement. For instance in tasks such as background subtraction, e.g., in IR spectroscopy, whereby bulk water signal is removed from a spectrum to highlight the signal of an analyte in solution. In the context of protein density measurement, experimental techniques that estimate protein density often assume that water is a medium of constant density [6]. The density of a weakly hydrated protein might be overestimated if water is treated purely as bulk though, as part of the "extra mass" observed is explained by water being on average denser around the protein.

ACKNOWLEDGMENTS

We wish to thank Durham HPC Hamilton for computational resources. We acknowledge the EPSRC and the SOFI2 CDT (grant EP/S023631/1) for financial support.

VI. AUTHOR DECLARATIONS

A. Conflict of Interest

The authors have no conflict of interest to disclose.

B. Author Contributions

Cameron C. W. McAllister: Conceptualization, Methodology, Investigation, Software, Formal Analysis, Visualization, Writing - Original Draft Preparation. **Lucas S. P. Rudden** Conceptualization, Resources. **Elizabeth H. C. Bromley** Conceptualization, Methodology, Supervision, Writing - Original Draft Preparation. **Matteo T. Degiacomi**, Conceptualization, Methodology, Software, Investigation, Supervision, Writing - Original Draft Preparation.

VII. DATA AVAILABILITY

Our method to calculate protein densities is available at www.github.com/Degiacomi-Lab/ProteinDensity.
The trained random forest regressors are available at www.github.com/Degiacomi-Lab/DensiTree. The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1] P. B. Chetri, H. Khan, and T. Tripathi, "Methods to determine the oligomeric structure of proteins," Advances in Protein Molecular and Structural Biology Methods , 49–76 (2022).
[2] S. Wheaton, R. M. Gelfand, and R. Gordon, "Probing the raman-active acoustic vibrations of nanoparticles with extraordinary spectral resolution," Nature Photonics **9**, 68–72 (2015).
[3] B. W. Matthews, "Solvent content of protein crystals," Journal of Molecular Biology **33**, 491–497 (1968).
[4] H. P. Erickson, "Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy," Biological Procedures Online **11**, 32 (2009).
[5] K. M. Andersson and S. Hovmöller, "The average atomic volume and density of proteins," Zeitschrift fur Kristallographie - New Crystal Structures **213**, 369–373 (1998).
[6] K. Gekko and H. Noguchi, "Compressibility of globular proteins in water at 25°c," Journal of Physical Chemistry **83**, 2706–2714 (1979).
[7] P. G. Squire and M. E. Himmel, "Hydrodynamics and protein hydration," Archives of Biochemistry and Biophysics **196**, 165–177 (1979).
[8] J. Greer and B. L. Bush, "Macromolecular shape and surface maps by solvent exclusion." Proceedings of the National Academy of Sciences **75**, 303–307 (1978).
[9] C. R. Chen and G. I. Makhatadze, "Proteinvolume: Calculating molecular van der waals and void volumes in proteins," BMC Bioinformatics **16**, 1–6 (2015).
[10] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," Journal of molecular biology **55** (1971), 10.1016/0022-2836(71)90324-X.
[11] T. J. Richmond, "Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect," Journal of Molecular Biology **178**, 63–89 (1984).
[12] H. Fischer, I. Polikarpov, and A. F. Craievich, "Average protein density is a molecular-weight-dependent function," Protein Science **13**, 2825–2828 (2004).
[13] M. L. Quillin and B. W. Matthews, "Accurate calculation of the density of proteins," Acta crystallographica. Section D, Biological crystallography **56**, 791–794 (2000).

[14]J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, "The packing density in proteins: standard radii and volumes," Journal of Molecular Biology **290**, 253–266 (1999).
[15]N. Q. Vinh, S. J. Allen, and K. W. Plaxco, "Dielectric spectroscopy of proteins as a quantitative experimental test of computational models of their low-frequency harmonic motions," Journal of the American Chemical Society **133**, 8942–8947 (2011).
[16]S. Martini, C. Bonechi, A. Foletti, and C. Rossi, "Water-protein interactions: The secret of protein dynamics," The Scientific World Journal **2013** (2013), 10.1155/2013/138916.
[17]L. Biedermannová and B. Schneider, "Hydration of proteins and nucleic acids: Advances in experiment and theory. a review," Biochimica et Biophysica Acta (BBA) - General Subjects **1860**, 1821–1835 (2016).
[18]M. Sanner, A. Olson, and J. Spehner, "Reduced surface: An efficient way to compute molecular surfaces," Biopolymers (1996), 10.1002/(SICI)1097-0282(199603)38:3.
[19]T. Can, C. I. Chen, and Y. F. Wang, "Efficient molecular surface generation using level-set methods," Journal of Molecular Graphics and Modelling **25**, 442–454 (2006).
[20]M. Hutt, T. Kulschewski, and J. Pleiss, "Molecular modelling of the mass density of single proteins," http://dx.doi.org/10.1080/07391102.2012.680031 **30**, 318–327 (2012).
[21]H. Zhao, P. H. Brown, and P. Schuck, "On the distribution of protein refractive index increments," Biophysical Journal **100**, 2309–2317 (2011).
[22]A. Goede and R. Preissner, "Voronoi cell: New method for allocation of space among atoms: Elimination of avoidable errors in calculation of atomic volume and density," Journal of Computational Chemistry (1997).
[23]M. Gerstein, J. Tsai, and M. Levitt, "The volume of atoms on the protein surface: Calculated from simulation, using voronoi polyhedra," J. Mol. Biol **249**, 955–966 (1995).
[24]T. K. Ho, "Random decision forests," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR **1**, 278–282 (1995).
[25]J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," nature **267**, 585–590 (1977).
[26]D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," Science **330**, 341–346 (2010).
[27]H. Granzier, M. Kellermayer, M. Helmes, and K. Trombitás, "Titin elasticity and mechanism of passive force development in rat cardiac myocytes probed by thin-filament extraction." Biophysical Journal **73**, 2043 (1997).
[28]C. A. Opitz, M. Kulke, M. C. Leake, C. Neagoe, H. Hinssen, R. J. Hajjar, and W. A. Linke, "Damped elastic recoil of the titin spring in myofibrils of human myocardium," Proceedings of the National Academy of Sciences of the United States of America **100**, 12688 (2003).
[29]M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, "Reversible unfolding of individual titin immunoglobulin domains by afm," Science **276**, 1109–1112 (1997).
[30]D. Svergun, S. Richard, M. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, "Protein hydration in solution: experimental observation by x-ray and neutron scattering," Proceedings of the National Academy of Sciences **95**, 2267–2272 (1998).
[31]F. Merzel and J. C. Smith, "Is the first hydration shell of lysozyme of higher density than bulk water?" Proceedings of the National Academy of Sciences **99**, 5378–5383 (2002).
[32]T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P. A. Bates, J. Fernandez-Recio, A. M. Bonvin, and Z. Weng, "Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2," Journal of molecular biology **427**, 3031–3041 (2015).
[33]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 **11**, 3696 (2015).
[34]T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An n·log(n) method for ewald sums in large systems," The Journal of Chemical Physics **98**, 10089–10092 (1993).
[35]G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," The Journal of chemical physics **126** (2007), 10.1063/1.2408420.
[36]M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," Journal of Applied Physics **52**, 7182–7190 (1981).
[37]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," Proteins **78**, 1950 (2010).
[38]J. B. Linse and J. S. Hub, "Scrutinizing the protein hydration shell from molecular dynamics simulations against consensus small-angle scattering data," Communications Chemistry 2023 6:1 **6**, 1–10 (2023).
[39]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 **585**, 357–362 (2020).
[40]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. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. 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, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods **17**, 261–272 (2020).
[41]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).
[42]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 **32**, 2319–2327 (2011).
[43]H. Liu, S. Xiang, H. Zhu, and L. Li, "The structural and dynamical properties of the hydration of snase based on a molecular dynamics simulation," Molecules **26** (2021), 10.3390/MOLECULES26175403.
[44]A. Bondi, "Van der waals volumes and radii," Journal of Physical Chemistry **68**, 441–451 (1964).
[45]S. Park and J. G. Saven, "Statistical and molecular dynamics studies of buried waters in globular proteins," Proteins: Structure, Function, and Bioinformatics **60**, 450–463 (2005).
[46]M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," KDD-96 Proceedings (1996).
[47]F. Pedregosa and et al., "Scikit-learn: Machine learning in python," Journal of Machine Learning Research **12**, 2825–2830 (2011).
[48]G. Shannon, C. R. Marples, R. D. Toofanny, and P. M. Williams, "Evolutionary drivers of protein shape," Scientific Reports 2019 9:1 **9**, 1–15 (2019).
[49]A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. lysozyme and insulin," Journal of molecular biology **79** (1973), 10.1016/0022-2836(73)90011-9.
[50]K. T. Chang and C. I. Weng, "The effect of an external magnetic field on the structure of liquid water using molecular dynamics simulation," Journal of Applied Physics **100** (2006), 10.1063/1.2335971.
[51]S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika **52**, 591–611 (1965).
[52]M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz, and F. Noé, "Pyemma 2: A software package for estimation, validation, and analysis of markov models," Journal of Chemical Theory and Computation **11**, 5525–5542 (2015).
[53]L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," Physical Review Letters **72**, 3634 (1994).
[54]G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé, "Identification of slow molecular order parameters for markov model construction," Journal of Chemical Physics **139**, 15102 (2013).
[55]S. P. Lloyd, "Least squares quantization in pcm," IEEE Transactions on Information Theory **28**, 129–137 (1982).

[56] J.-H. Prinz, B. Keller, and F. Noé, "Probing molecular kinetics with markov models: metastable states, transition pathways and spectroscopic observables," Physical Chemistry Chemical Physics **13**, 16912–16927 (2011).
[57] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid monte carlo," Journal of Computational Physics **151**, 146–168 (1999).
[58] S. Röblitz and M. Weber, "Fuzzy spectral clustering by pcca+: Application to markov state models and data classification," Advances in Data Analysis and Classification **7**, 147–179 (2013).
[59] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "Mdtraj: A modern open library for the analysis of molecular dynamics trajectories," Biophysical Journal **109**, 1528 – 1532 (2015).
[60] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," Biopolymers: Original Research on Biomolecules **22**, 2577–2637 (1983).
[61] T. J. McIntosh and S. A. Simon, "Area per molecule and distribution of water in fully hydrated dilauroylphosphatidylethanolamine bilayers," Biochemistry **25**, 4948–4952 (1986).
[62] D. Laage, T. Elsaesser, and J. T. Hynes, "Water dynamics in the hydration shells of biomolecules," Chemical Reviews **117**, 10694–10725 (2017).