STRUCTURAL BIOLOGY
ISSN 2059-7983

SCOMAP-XD: atomistic deuterium contrast matching for small-angle neutron scattering in biology

Alan Hicks,[a] Paul Abraham,[b] Wellington Leite,[c] Qiu Zhang,[c] Kevin L. Weiss,[c] Hugh O'Neill,[c] Loukas Petridis[a]* and Jeremy C. Smith[a,d]*

[a]UT/ORNL Center for Molecular Biophysics, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, [b]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, [c]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, and [d]Department of Biochemistry and Cellular and Molecular Biology, The University of Tennessee, Knoxville, 309 Ken and Blaire Mossman Building, 1311 Cumberland Avenue, Knoxville, TN 37996, USA. *Correspondence e-mail: loukisp@gmail.com, smithjc@ornl.gov

Received 17 January 2023
Accepted 27 March 2023

Edited by P. Langan, Institut Laue–Langevin, France

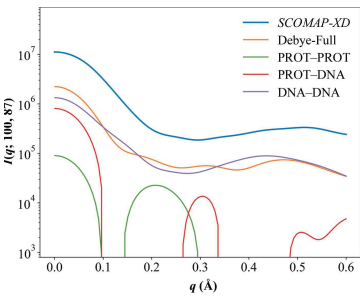
Keywords: small-angle neutron scattering; contrast matching; software; biomolecular structure.

Supporting information: this article has supporting information at journals.iucr.org/d

The contrast-variation method in small-angle neutron scattering (SANS) is a uniquely powerful technique for determining the structure of individual components in biomolecular systems containing regions of different neutron scattering length density ρ. By altering the ρ of the target solute and the solvent through judicious incorporation of deuterium, the scattering of desired solute features can be highlighted. Most contrast-variation methods focus on highlighting specific bulk solute elements, but not on how the scattering at specific scattering vectors q, which are associated with specific structural distances, changes with contrast. Indeed, many systems exhibit q-dependent contrast effects. Here, a method is presented for calculating both bulk contrast-match points and q-dependent contrast using 3D models with explicit solute and solvent atoms and *SASSENA*, an explicit-atom SANS calculator. The method calculates the bulk contrast-match points within 2.4% solvent D_2O accuracy for test protein–nucleic acid and lipid nanodisc systems. The method incorporates a general model for the incorporation of deuterium at non-exchangeable sites that was derived by performing mass spectrometry on green fluorescent protein. The method also decomposes the scattering profile into its component parts and identifies structural features that change with contrast. The method is readily applicable to a variety of systems, will expand the understanding of q-dependent contrast matching and will aid in the optimization of next-generation neutron scattering experiments.

1. Introduction

Small-angle scattering (SAS) of X-rays (SAXS) and neutrons (SANS) provides low-resolution information on molecular assemblies. These techniques have been widely adopted by structural biologists to understand the sizes and shapes of proteins, nucleic acids, lipids and plant biomass in solution as well as their complex molecular assemblies (Jacrot, 1976; Timmins & Zaccai, 1988; Engelman & Moore, 1975). For neutrons, the contrast-variation method is a unique and fundamental tool for highlighting individual components in multicomponent systems (Engelman & Moore, 1975; Krueger, 2022), effectively separating the scattering from a multi-component complex into contributions from its components through the use of hydrogen–deuterium substitution in the solute and/or the solvent. By tuning the contrast, *i.e.* the difference in the scattering length densities between a given solute component and the solvent, through changing the H_2O to D_2O ratio of the solvent, the scattering of the solute component can be 'matched'. At this so-called contrast-match



point the scattering of the solute component is severely attenuated, thereby capturing only the scattering of another, target solute component for which one wishes to determine the structure. The calculation of these bulk 'match points' is a fundamental step for efficient SANS experimental design and system preparation (Jeffries *et al.*, 2016), but does not furnish information on how scattering on specific length scales changes with contrast. In this study, we develop a method that uses structure-based SANS calculators to investigate this phenomenon explicitly.

Deuterium has a much larger coherent scattering length than its isotope hydrogen: 6.67 versus −3.74 fm. The incorporation of deuterium into a solute is an important aspect and tool for the contrast-variation method in SANS. Three methods are known by which deuterium can be incorporated into biological samples. Hydrogen–deuterium (H–D) exchange of labile H atoms bound to nitrogen, oxygen, sulfur and phosphorus with D_2O in the solution is the most basic of these methods (Bai *et al.*, 1993; Best & Vendruscolo, 2006). When the kinetic rate of H–D exchange is slower than the rate of local unfolding of the protein, the protection factor, *i.e.* the degree to which any exchangeable hydrogen is likely to exchange, is the ratio of the intrinsic exchange rate of the amino-acid sequence from a tripeptide in solution to the measured exchange rate (Bai *et al.*, 1993; Best & Vendruscolo, 2006). A second method is by the perdeuteration of selected amino acids (Laux *et al.*, 2008) or whole domains (Sonntag *et al.*, 2017). In one example of this, Sonntag and coworkers performed an experiment on the TIA1–RNA binding system to understand the role of the three RNA-recognition motif (RRM) domains in binding to a target RNA (Sonntag *et al.*, 2017). They separately perdeuterated different domains of TIA1, the first RRM domain (R1) and the second and third RRM domains (R2R3), and then spliced the perdeuterated domains to the corresponding protiated domains. Finally, the non-exchangeable H atoms can be substituted for deuterium through growth of microorganisms in D_2O medium (Lederer *et al.*, 1986; Leiting *et al.*, 1998; Weiss *et al.*, 2021). Previous studies of the incorporation of deuterium into proteins by growth in D_2O medium has shown a quadratic dependence of the fraction of deuteration at non-exchangeable sites on the fraction of D_2O in the solvent (Moore, 1977; Perkins, 1981; Lederer *et al.*, 1986; Leiting *et al.*, 1998). The amount of incorporated deuterium also depends on the addition of deuterated glucose or glycerol to the growth medium (Perkins, 1981). In general, bacteria and yeast prefer H_2O during growth, so it is necessary to adapt these organisms to D_2O growth conditions (Jeffries *et al.*, 2016; Weiss *et al.*, 2021).

The complexity of deuteration schemes underlines why the ability to calculate contrast-match points for a range of growth and solvent conditions would greatly improve the efficiency of experimental SANS design. Contrast-match points can be determined experimentally by performing a contrast-variation series, in which a series of SANS experiments are performed in solvent with different H_2O:D_2O ratios to detect where the scattering disappears or the forward scattering intensity $I(0)$ is zero (Jacrot, 1976; Jeffries *et al.*, 2016). There are numerous examples of this approach, but we highlight two studies used as test cases in the work that we report here. In one study, extensive SAXS and contrast-matched SANS experiments were performed to understand a model of the disordered binding of KorA/KorB to DNA operons (Hyde *et al.*, 2017). The authors expressed KorA in 0%, 46% and 87% D_2O growth medium. When 46% D_2O growth medium was used, KorA was experimentally determined to match out in 65% D_2O solvent, the same as for DNA. In 87% D_2O growth medium KorA matches out in 100% D_2O solvent. The other application of contrast matching and solution SANS involves the use of lipid nanodiscs, which are lipid bilayers constrained using membrane structural proteins, to study membrane proteins. Several studies have developed techniques to create lipid nanodiscs with various levels of deuteration (Maric *et al.*, 2014) but, as with protein systems, growth and preparation is difficult. The authors designed a deuterated lipid nanodisc from *Escherichia coli* cell cultures that was matched out in 100% solvent D_2O, thereby only observing proteins embedded in the nanodiscs. The lipid nanodiscs consisted of 93% deuterated palmitoleic acid and cycloproponated palmitic acid with the PC head groups deuterated at 78% (Maric *et al.*, 2014).

Theoretical calculations have been developed to determine the contrast-match points from the elemental composition or sequence of proteins and nucleic acids. Two popular methods for calculating contrast-match points that have been used for experimental design are *MULCh* (Whitten *et al.*, 2008) and *SASSIE* (Sarachan *et al.*, 2013). Originally, *MULCh* was used to determine the contrast-match points of two proteins in the Kin–SDA complex (Whitten *et al.*, 2007). Shortly afterwards, *SASSIE* helped to design the contrast series for the Skp–OmpA protein–protein complex and the Cre–LoxP protein–DNA complex (Sarachan *et al.*, 2013). These calculators focus on the bulk scattering match points, *i.e.* the match points where the scattering of whole molecules is removed. Both *SASSIE* and *MULCh* use the total average fraction of incorporated deuterium over the whole molecule, *i.e.* 'bulk' deuteration, to calculate the overall scattering length densities of biomolecules, but do not use 3D structural information. Strictly speaking, this approach assumes that the distribution of hydrogen–deuterium exchange is uniform throughout the sample. However, the radius of gyration and the scattering features at higher q vectors can depend on the contrast (Ibel & Stuhrmann, 1975). Furthermore, geometric parameters such as the radius of gyration cannot easily be determined if the distribution of H–D exchange throughout the sample is inhomogeneous. Discrepancies at higher q are also possible with the homogenous approximation (Witz, 1983). Structure-based explicit deuteration models coupled with structure-based SANS calculators provide a promising method for overcoming these challenges and for improving our understanding of q-dependent contrast effects.

The use of 3D structure-based SANS calculators, where one inputs the 3D coordinates of a biomolecule and the program outputs a SANS profile, has been a key advance in the analysis and interpretation of neutron scattering data. All-atom

structures and molecular-dynamics simulations on nanosecond to microsecond timescales have never been more accessible, due in part to *AlphaFold*2 (Jumper *et al.*, 2021; Mirdita *et al.*, 2022) structure predictors and GPU acceleration. There are several SANS calculators in use today (Svergun *et al.*, 1998; Grudinin *et al.*, 2017), but only a few consider explicit solvent (Merzel & Smith, 2002*a*; Park *et al.*, 2009; Lindner & Smith, 2012; Chen *et al.*, 2019). Implicit solvent SANS predictors such as *Pepsi-SANS* (Grudinin *et al.*, 2017) model the hydration shell by relying on a free parameter during fitting to the experimental data to account for the contrast in the hydration shell. These predictors do not capture explicit solvent scattering effects, such as H–D exchange between water molecules leading to HDO scattering in mixed solvents (Max & Chapados, 2002) or 'transparent water' scattering at 36% D_2O concentration (Powles *et al.*, 1972).

Accounting explicitly for the solvent scattering is important in part because of the hydration shell, *i.e.* the shell of water surrounding the solute, which can have a scattering length density ∼10% different from the bulk (Svergun *et al.*, 1998). It is well established that the contribution of the hydration shell has a large impact on structural modeling in small-angle X-ray and neutron scattering, particularly on the shapes of the individual particles or their intermolecular distances in a complex (Kim & Gabel, 2015). Correspondingly, explicit solvent calculators such as *SWAXS* (Chen & Hub, 2014), *SASSIM* (Merzel & Smith, 2002*a*) and its successor *SASSENA* (Lindner & Smith, 2012) have demonstrated an improvement over the implicit models (Chen & Hub, 2014). Using *SASSIM*, the authors identified key biophysical factors contributing to the increased solvent density in the hydration shell compared with bulk water (Merzel & Smith, 2002*b*).

Similar to the bulk contrast-matching calculators above, many 3D SANS calculators by default implement homogenous deuteration of exchangeable and non-exchangeable H atoms to handle the contrast. A recent study using a structural model to describe H–D exchange coupled with a kinetic model to determine the fraction of exchanged hydrogens over a period of time showed that understanding specific deuteration versus bulk deuteration is necessary to interpret scattering at higher q-values (Pedersen *et al.*, 2019). Explicitly modeled deuteration studies are necessary to identify how the explicit deuteration affects the contrast-match point and the q-dependency.

In this study, we develop and present *SCOMAP-XD*, a computational tool to atomistically explicitly deuterate biomolecular 3D structures and to investigate q-dependent contrast-match points using a combination of empirical structural models and *SASSENA*. In *SCOMAP-XD* explicitly deuterated structural models are used to predict the effect of deuteration on neutron scattering profiles and to estimate contrast-match points by calculating SANS profiles.

We first introduce the theory of contrast variation and neutron scattering that underpins our method. We proceed to discuss the methodology and the code itself. We apply *SCOMAP-XD* to a variety of different biomolecular test systems, such as human serum albumin (HSA), the KorA–DNA complex and a lipid nanodisc. We validate the calculation of SANS contrast-match points from calculated SANS profiles, the importance of the explicit deuteration and the selection of models for the solvent. In several systems, we observe clear q-dependent contrast effects. For a test protein–DNA system these effects are correlated with specific structural features by calculating the component scattering profiles using the Debye formula. As neutron sources and SANS instrumentation continually improve and the importance of large complexes in biology becomes more recognized, *SCOMAP-XD* can easily be usefully applied to a wide variety of biomolecular and soft-matter systems across multiple length scales.

2. Methods

2.1. Theory

We first present a brief description of scattering theory necessary to understand the proposed method. The scattering intensity $I(q)$ of the solute at a scattering vector q depends on the number density of scatterers in the sample (n), the contrast ($\Delta\rho$), the volume ($V_{\rm solute}$) of the solute, the form factor of the solute [$P(q)$] and the structure factor [$S(q)$] as given in equation (1) (Jacrot, 1976; Whitten *et al.*, 2008),

$$I(q) = n(V_{\rm solute}\Delta\rho)^2 P(q)S(q). \quad (1)$$

The contrast $\Delta\rho$ is the difference in the scattering length densities, *i.e.* the sum of the atomic neutron scattering lengths (b) over the volume, of the solute, $\sum b_i/V_{\rm solute}$, from the solvent, $\sum b_j/V_{\rm solvent}$,

$$\Delta\rho = \frac{\Sigma b_i}{V_{\rm solute}} - \frac{\Sigma b_j}{V_{\rm solvent}}. \quad (2)$$

$P(q)$ describes the intramolecular scattering from one particle, while $S(q)$ describes intermolecular structural correlations. In a dilute, monodisperse sample $S(q)$ is unity, as there is assumed to be no scattering between individual monomers, nor aggregation. Furthermore $P(0) = 1$, thus demonstrating that the forward scattering intensity is proportional to the squared contrast. By taking the square root of the forward scattering intensity $I(0)$, one obtains a linear function with respect to the contrast. The contrast-match point is defined as the contrast when $I(0)^{1/2} = 0$.

In the case of a bimolecular complex, the scattering intensity can be broken down into individual contributions as follows (Sarachan *et al.*, 2013),

$$I(q) = \Delta\rho_1 I_{11}(q) + \Delta\rho_1\Delta\rho_2 I_{12}(q) + \Delta\rho_2 I_{22}(q), \quad (3)$$

where $\Delta\rho_1$ and $\Delta\rho_2$ are the contrasts of each component calculated from equation (2) and $I_{ij}(q)$ is the intramolecular scattering when $i = j$ and the intermolecular scattering when $i \neq j$. It is clear from equation (3) that when one component is contrast-matched out, *i.e.* when $\Delta\rho_i = 0$, only the scattering of the other component is observed. This formula can be expanded to multiple-component complexes, so by tuning the relative contrast we can resolve the shape and the structure of

a single component in a multiple-component complex. A fundamental consideration when designing SANS contrast experiments of complexes is the difference between the contrast of the individual component and that of the complex (following equation 1). If the contrast at which the experiment is performed is too close to the contrast of the complex, the intensity may be too low to identify the scattering of the target.

2.2. *Scattering Contrast Match Points with Explicit-atom Deuteration* (*SCOMAP-XD*)

SCOMAP-XD is divided into three sections (Fig. 1). The first section is a Python script which takes a solvated, preferably energy-minimized, structure as input along with two experimental D_2O conditions: the percentage of D_2O in the solvent in the SANS experiments (sD_2O) and, where appropriate, the percentage of D_2O used in the growth medium (gD_2O) during recombinant protein expression. The outputs are a set of deuterated PDB files as templates for *SASSENA*. In Section 2, *SASSENA* uses these template PDB files and a DCD-formatted MD trajectory or single snapshot to calculate the neutron scattering profiles. The third section is the calculation of contrast-match points written in Python. Further analysis can be performed on the SANS results if desired.

We now describe the flow of the program and the implementation of the deuteration models. The Python scripts for deuteration, the necessary *SASSENA* input files, submission scripts for analysis and match-point prediction scripts are available at https://github.com/achicks15/SCOMAP-XD.

2.2.1. Modeling the explicit deuteration of biomolecular structures. The deuteration module handles the substitution of hydrogen for deuterium at both exchangeable and non-exchangeable sites for the solute and the solvent. The inputs of the Python script are a solvated PDB file, the percentage of D_2O [%(*v*/*v*)] in the solvent used for SANS measurements and the percentage of D_2O [%(*v*/*v*)] in the growth medium used to grow *E. coli* cultures for the production of deuterated protein. If lipids are present in the solute, the fraction of deuteration can be individually added for each acyl chain and the head group. The user can also provide a set of solvent and/or growth D_2O conditions to calculate the set of deuterated PDB files necessary for a contrast-match point calculation. The code will parse the topologies using *mdtraj* (McGibbon *et al.*, 2015) to read the PDB files and determine the macromolecular content of the solute. The H atoms in each component are categorized into exchangeable (N—H, O—H, S—H and P—H) and non-exchangeable (C—H), and are randomly chosen for deuteration according to modeled weights for each hydrogen type. Exchangeable H atoms are weighted according to their protection factor (Bai *et al.*, 1993; see below), while non-exchangeable H atoms have equal weights. The names and elements of selected H atoms are then converted to deuterium ('D') in the PDB file for the *SASSENA* template.

The exchangeable H atoms are weighted by the protection factor calculated using the model developed by Best & Vendruscolo (2006). The protection factor is the ratio of the NMR hydrogen–deuterium exchange rate k_{ex} and the intrinsic sequence-dependent exchange rate k_{int}. The higher the protection factor, the less likely the hydrogen is to exchange with the solvent. The model predicts the log of the protection factor from the number of heavy-atom contacts within 6.5 Å and the number of hydrogen bonds. We handle calculation of the hydrogen bonds and heavy-atom contacts with the *mdtraj* Python module. The protection factors relate to the free energy of local unfolding, where the solute is expected to partially unfold and then exchange with the solvent. Thus, we weigh each exchangeable H atom by taking the inverse of its protection factor and normalize to the sum of the weights. This

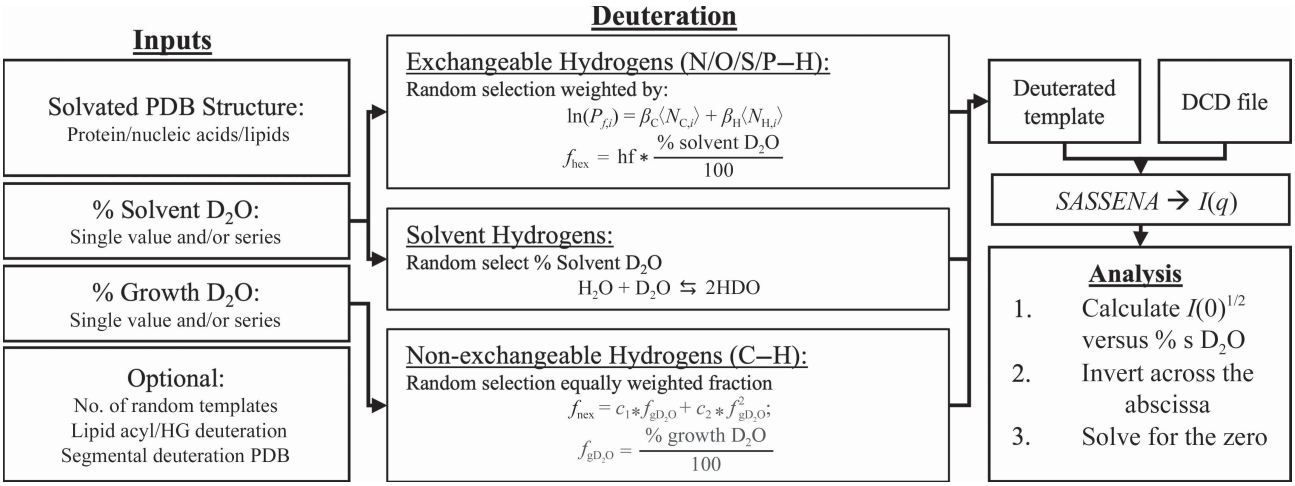


Figure 1
Diagram of the *SCOMAP-XD* workflow: the inputs, deuteration step, outputs for *SASSENA* and calculation of the contrast-match point analysis. The specific models are displayed in the deuteration section and the arrows from the inputs highlight how the different D_2O inputs are used in the deuteration step. We describe the deuteration models for exchangeable and non-exchangeable H atoms in Section 2.2.1 and for solvent H atoms in Section 2.2.2. A brief introduction to *SASSENA* is given in Section 2.2.3. Section 2.2.4 contains a description of the analysis to calculate the contrast-match point.

model was originally designed for backbone peptide groups of proteins. We assume that these concepts are adaptable to all exchangeable H atoms, so we apply the model as such. For complexes, the protection factor is calculated for the whole complex and then averaged between the single molecule and complex (Zhang *et al.*, 2011). The total fraction of exchangeable H atoms selected for deuteration is set by the percentage solvent D_2O (% sD_2O), *i.e.* if sD_2O is 20% only 20% of the exchangeable H atoms can be selected. The total number of exchangeable H atoms is scaled by an exchange factor, *hf*, between 0 and 1 to account for incomplete exchange with the solvent. In proteins, the exchange factor is typically set to 0.95 (Perkins, 1986), whereas for nucleic acids and lipid head groups it is 1 (Jacrot, 1976).

To derive a general model for the incorporation of deuterium at non-exchangeable sites, we performed mass spectrometry on green fluorescent protein (GFP) grown in 0%, 20%, 40%, 70% and 100% D_2O and built an empirical model from these data. Previous experiments have shown a quadratic dependence of the total fraction of deuterium incorporated into the non-exchangeable sites of the solute on the D_2O in the growth medium (Lederer *et al.*, 1986; Leiting *et al.*, 1998). GFP was prepared in 0%, 20%, 40%, 70% and 100% D_2O growth medium and then processed for mass spectrometry (Section 2.3). Mass spectrometry will detect the difference in mass of a peptide fragment due to deuterium incorporation. The total fraction of deuterated non-exchangeable hydrogen is the ratio of the difference between the mass of the fragment grown in D_2O and the expected protiated mass of the peptide over the total mass of the hydrogens in the peptide. The mass-spectrometric data for each peptide fragment of GFP were fitted by a quadratic function,

$$f_{\rm nex} = c_1 f_{{\rm gD_2O}} + c_2 f_{{\rm gD_2O}}^2, \tag{4}$$

where $f_{\rm nex}$ is the fraction of non-exchangeable H atoms that are deuterated and $f_{\rm gD_2O}$ is the fraction of D_2O in the growth medium, between 0 and 1. The c_1 and c_2 parameters from each fit were averaged together for the final weights.

For deuteration, we select a subset, given by $f_{\rm nex}$, of the non-exchangeable H atoms in the structure at random with equation (4). During selection, every non-exchangeable H atom is given equal weight. To test the model, we deuterated GFP (PDB entry 1gfl; Yang *et al.*, 1996) at 0%, 20%, 40%, 70% and 100% gD_2O. For each growth condition, we created 100 random deuterium distributions in the non-exchangeable H atoms throughout the structure. We do not consider the addition of deuterated carbon sources during growth in the model. Also, currently, nucleic acids are not deuterated at non-exchangeable sites. Lipids can be deuterated using a flag in the command line to individually deuterate each acyl chain and the head groups. In the case of perdeuteration, where all non-exchangeable sites are deuterated, or segmental deuteration of the solute at non-exchangeable sites, a template PDB file should be provided with the occupancy column for each non-exchangeable H atom set to 1.

2.2.2. Modeling solvent-specific effects using all-atom explicit solvent. In the model, explicit waters are added to the PDB structure to account for solvent scattering. *SASSENA* also calculates the background scattering length density with the explicit solvent. Three models were tested for water deuteration, where the input parameter % sD_2O determines the total number of H atoms to substitute: random deuteration of H atoms, keeping D_2O whole and an exchange model. Random deuteration treats each hydrogen in water equally. Keeping D_2O whole is an unphysical model which deuterates both H atoms on one water molecule accounting for the total incorporation of deuterium into the solvent. The third, a realistic model we call ExH, counts the molar fraction of HDO according to the equilibrium exchange equation $H_2O + D_2O \leftrightarrow 2HDO$ (Max & Chapados, 2002). The equilibrium constant was set to $K_{\rm eq} = 3.73$ and the corresponding set of nonlinear equations (Max & Chapados, 2002) was solved using the *fsolve* package in Python at any fraction of sD_2O to find the number of molecules that would be HDO or D_2O.

2.2.3. Calculating scattering profiles from 3D structures. *SASSENA* is a highly scalable, all-atom scattering calculator for SAXS or SANS. It calculates the coherent scattering intensity $I(q)$ from the scattering amplitudes of all atoms in the designated system. The scattering intensity and amplitude for SANS is

$$I(q) = A^*(q)A(q);\ A(q) = \sum_{n=0}^{N} b_n \exp(-i\mathbf{r}_n\mathbf{q}), \tag{5}$$

where b_n is the coherent neutron scattering length of atom n, $\mathbf{r}_n$ is the vector coordinate of atom n and $\mathbf{q}$ is the scattering vector, where bold text denotes a vector. The magnitudes of the scattering vectors are distributed uniformly between 0 and a provided maximal value, while the vectors associated with each magnitude are randomly distributed over the unit sphere and are averaged over. For each magnitude, the user can select a number of vectors to average to sample the full solid angle, typically 1000 vectors. *SASSENA* uses Babinet's principle to handle the contrast via subtraction of the bulk-solvent scattering length density $\bar{b}_0$ with the solvent-excluded volume of atom n, V_n,

$$b_n \rightarrow b_n - V_n\kappa_n\bar{b}_0, \tag{6}$$

where κ_n is a scaling factor associated with the packing of the atom, whether it is part of the solute or the solvent. In this case, atoms in the solute are scaled by 1.0, whereas atoms in the solvent are scaled by 1.52. These values were found to give the lowest χ^2 to the experimental contrast-match points and known scattering length densities of H_2O and D_2O. The volume V_n is the crystallographic solvent-excluded volume. The bulk-scattering length density is set at the start of the calculation; it can either be a user-defined scattering length density of the solvent or *SASSENA* can calculate the background as the sum of the atomic scattering lengths divided by the sum of the atomic volumes. *SASSENA* will average the scattering profiles over a series of snapshots, such as from a molecular-dynamics (MD) trajectory.

2.2.4. Analysis and solving the contrast-match point. As stated in Section 2.1, the contrast is proportional to the square root of the forward scattering intensity $I(0)^{1/2}$. *SASSENA* calculates $I(0)$ for each SANS profile for a series of different % sD_2O values, typically 5–7 values from 0% to 100% sD_2O. There will be an inflection point in the plot of $I(0)^{1/2}$ versus % sD_2O because the contrast shifts from negative to positive for most macromolecules (Timmins & Zaccai, 1988). Therefore, the data need to be inverted across the x axis from the sign degeneracy in $I(0)^{1/2}$. To find the inflection point, the calculated contrast series is interpolated via cubic spline and evaluated over a series of % sD_2O values from 0% to 100 in 1% increments using the *scipy* interpolate module. The minimum of the interpolated curve is the inflection point, where the first derivative is zero. The two closest points to the minima are selected from the original data and all the points after the first positive-sloped value are inverted. Only points less than 100% D_2O are selected in the fits, as it is difficult to truly determine the inversion points when the match point is close to 100% sD_2O. The same method is applied to the inverse problem to solve for the growth conditions of a protein to match out at a specific % sD_2O.

2.3. Molecular-dynamics simulations and system preparation

Supplementary Table S1 lists the test systems that we used to validate the model. Almost all systems were taken from the PDB and were prepared using *charmm-gui* (Jo *et al.*, 2008; Lee *et al.*, 2016), unless specified otherwise. Human serum albumin (HSA) was retrieved from SASBDB (Kikhney *et al.*, 2020) entry SASDAA6 (Franke *et al.*, 2015). The maltose-binding protein (PDB entry 1omp) system has a His tag added (Laux *et al.*, 2008) using *AlphaFold*2 (Jumper *et al.*, 2021; Mirdita *et al.*, 2022). The lipid nanodisc structure was prepared according to Maric *et al.* (2014) with the 97A MSP1D1 and PYPC lipids only in *charmm-gui* (Jo *et al.*, 2008; Lee *et al.*, 2016; Qi *et al.*, 2019). The cyclopropanated lipids were not included as they were not available. The TIA1 system was built using *AlphaFold*2 *ColabFold* (Mirdita *et al.*, 2022) by providing the TIA1 sequence with the appropriate LPQTG mutation in the first linker region (Sonntag *et al.*, 2017) with PDB entries 5o2v (Sonntag *et al.*, 2017) and 2mjn (Wang *et al.*, 2014) as templates.

The systems were subjected to molecular-dynamics (MD) simulation to provide realistic conformational ensembles for the macromolecules and solvent molecules. Each system was simulated for 10 ns according to the *charmm-gui* (Lee *et al.*, 2016) protocols in *GROMACS* 2018.8 (Van Der Spoel *et al.*, 2005; Abraham *et al.*, 2015) using the Charmm36m force field (Huang *et al.*, 2017) and TIP3P waters (Jorgensen *et al.*, 1983). Systems were first energy-minimized using the steepest-descent algorithm for 5000 steps. Temperature equilibration to the target temperature, typically 298 K, was performed with the Nosé–Hoover thermostat (Nosé, 1984; Hoover, 1985; Martyna *et al.*, 1992) with a coupling time of 1 ps over 125 ps with a 1 fs time step. Subsequently, ten steps of 1 ns simulation in the NPT ensemble were performed with a 2 fs leapfrog integration step. The temperature was maintained at 298 K with a 1 ps coupling constant using the same thermostat as the equilibration. The pressure was maintained at 1 atm with the Parrinello–Rahman barostat (Parrinello & Rahman, 1981). All simulations implemented the Verlet cutoff scheme with a switch cutoff at 10 Å and a hard cutoff at 12 Å for the short-range Lennard–Jones and electrostatic forces. Long-range electrostatics were handled with the particle mesh Ewald method (Darden *et al.*, 1993; Essmann *et al.*, 1995). Bonded H atoms were constrained with the *LINCS* algorithm (Hess *et al.*, 1997; Hess, 2008).

Debye decomposition was performed by calculating the pairwise distances between all atoms in the protein domains, DNA domains and between the two domains using *mdtraj*. The distances were averaged over the full MD trajectory from PDB entry 5cm3. The scattering intensity was then calculated using the formula

$$I(q) = \sum_{i,j}^{N} b_i b_j \frac{\sin(q r_{ij})}{q r_{ij}}, \tag{7}$$

where $b_{i,j}$ is the scattering length of atom i or j re-adjusted for the excluded volume and contrast as in *SASSENA* (see Section 2.2.3), q is the magnitude of the scattering vector spaced logarithmically from 0.001 to 0.6 $Å^{-1}$ and r_{ij} is the distance between atoms i and j. The template PDB entries prepared for *SASSENA* were used to define the deuterated atoms of the protein and DNA for the different contrasts.

The root-mean-square error (RMSE) and χ^2 for comparing the calculated contrast-match points with experimental results were calculated as

$$\chi^2 = \sum_{i}^{N} \frac{(\text{Experiment} - \text{Calculated})^2}{\text{Experiment}},$$
$$\text{RMSE} = \left(\frac{1}{N}\sum_{i}(\text{Experiment} - \text{Calculated})^2\right)^{1/2}, \tag{8}$$

where Experiment is the experimental contrast-match point and Calculated is the match point calculated using *SCOMAP-XD*. The R^2 correlation coefficient was determined using the OLS module from the *statsmodels* package in Python. Structures were visualized and rendered using *ChimeraX* (Pettersen *et al.*, 2021). All data were plotted using the Python modules *matplotlib*, *pandas* and *seaborn*. The scripts for the *SCOMAP-XD* workflow, including deuteration and analysis, are available at https://github.com/achicks15/SCOMAP-XD.

2.4. Protein production and characterization

The expression and purification of *Aequorea victoria* green fluorescent protein and mass-spectrometric analysis have been described previously (Weiss *et al.*, 2021).

Human serum albumin (HSA) solutions were prepared by dissolving lyophilized powder (Sigma–Aldrich, catalog No. A3782) in phosphate-buffered saline (PBS) consisting of 10 m*M* sodium phosphate, 140 m*M* NaCl, 3 m*M* KCl pH 7.4 (Millipore, catalog No. 524650) using either H_2O or D_2O as a solvent. Once dissolved, the solutions were passed through a 0.2 µm centrifugal filter (VWR, catalog No. 82031-358) and

purified using an ÄKTA Go protein-purification system (Cytiva) by size-exclusion chromatography on a Superdex 200 Increase 10/300 GL column (Cytiva, catalog No. 28990944) that had been pre-equilibrated with PBS in either H_2O or D_2O. The UV absorbance at 280 nm of the major peak fractions was measured using a NanoDrop One spectrophotometer (Thermo Scientific) and concentrations were determined using a mass extinction coefficient ($\varepsilon_{1\%}$) of 5.30 g^{-1} l cm^{-1}. The peak fractions from the H_2O-based and D_2O-based purifications were combined (~3–4 mg ml^{-1}), mixed in varying D_2O/H_2O ratios and centrifuged at 16 100*g* for 15 min before being carried on for small-angle neutron scattering measurements.

2.5. Small-angle neutron scattering

Small-angle neutron scattering (SANS) measurements were performed using the Bio-SANS instrument at the High Flux Isotope Reactor in Oak Ridge National Laboratory (Heller *et al.*, 2014). The configuration of the dual detector system with four guides was set as follows: the main detector was at 7 m from the sample position and the wing detector was at a fixed radius of 1.13 m from the sample and was rotated to an angle of 3.2° from the direct beam. In this configuration, a q range of $0.007 < q < 0.9$ Å^{-1} was obtained with a relative wavelength spread ($\Delta\lambda/\lambda$) of 13.2% and a neutron wavelength of 6 Å ($q = 4\pi\sin\theta/\lambda$, where 2θ is the scattering angle and λ is the neutron wavelength). The data were corrected for instrument background, detector sensitivity and instrument geometry by the facility data-reduction software *drtsans* using a Python script wrapper (Heller *et al.*, 2022). All SANS measurements were performed in cylindrical quartz cuvettes of 1 mm path length (Hellma, Müllheim, Germany) at 20°C.

3. Results

3.1. Empirical model parametrization and the effect of deuteration models on SANS

We begin by developing empirical models for explicit deuteration, as shown in Fig. 2. An example of the exchangeable hydrogen model is shown in Fig. 2(*a*) for GFP (PDB entry 1gfl) in 100% D_2O solvent, where the coloring is the $\ln(P_f)$ value. To determine the incorporation of deuterium at non-exchangeable sites, we performed mass spectrometry on GFP grown in 0%, 20%, 40%, 70%, 85% and 100% D_2O, as shown in Fig. 2(*b*). We fitted the percentage deuteration calculated from mass spectrometry at each growth percentage with a quadratic function (equation 4) for six peptide fragments. The sequences and structural positions of the fragments are shown in Supplementary Fig. S1. The c_1 and c_2 parameters from the fits in equation (6) were found to be 0.40 ± 0.08 and 0.46 ± 0.09, respectively. The model has excellent agreement with experiment: the χ^2 between the experimental results and model prediction of deuterated GFP segments after averaging over 100 randomizations of the non-exchangeable hydrogen deuteration pattern is between 0.752% and 1.62%.

We proceeded to determine how the variation in the solute deuteration patterns from the random selection process influenced the SANS profiles (Fig. 3). For this, we calculated

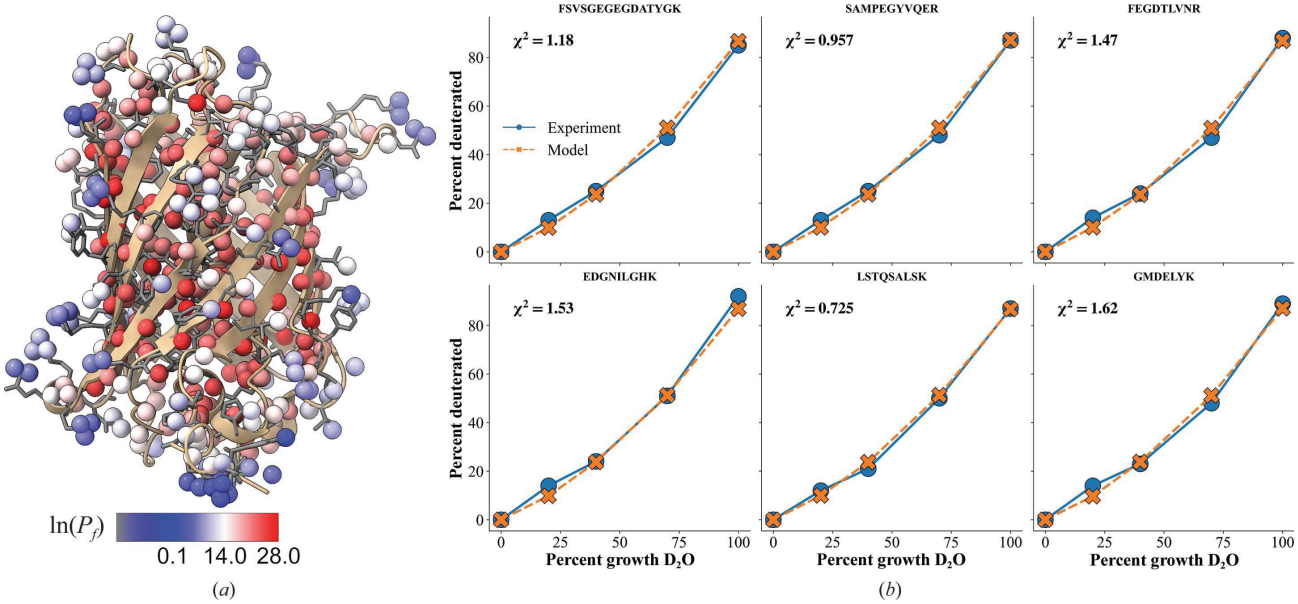


Figure 2
Exchangeable and non-exchangeable deuteration models. (*a*) GFP (PDB entry 1gfl) with exchangeable H atoms colorized by $\ln(P_f)$ calculated from the model of Best & Vendruscolo (2006). The red-colored H atoms in the core of GFP are more protected, whereas H atoms on the exterior, colored blue, are more likely to exchange. (*b*) Comparing the experimental percentage deuteration from mass spectroscopy (blue circles) for segments of GFP at different growth conditions with the average of 100 random non-exchangeable hydrogen deuteration patterns (orange crosses). The χ^2 of the fits are shown in their respective plots.

the SANS profiles over 100 different variations of GFP deuteration patterns for two solvent conditions, 0 and 100% solvent D_2O (sD_2O), and two growth conditions, 0 and 40% growth D_2O (gD_2O), in Fig. 3(*a*). The forward scattering intensity should be invariant as the scattering length density is the same averaged over the whole molecule regardless of how the deuterium is distributed. To verify that our selection of deuteration sites is working according to theory, we calculated the coefficient of variation (CoV), the ratio of the standard deviation to the mean intensity, at every q value (Fig. 3*c*). Indeed, the CoV value at $q = 0$ is zero for all curves. In general, there are large differences in the forward scattering intensity as the contrast changes. The scattering intensity loss from 0% to 100% sD_2O and the intensity gain upon adding non-exchangeable deuteration using 40% gD_2O are expected. In 100% sD_2O and 40% gD_2O, the intensity drops relative to the

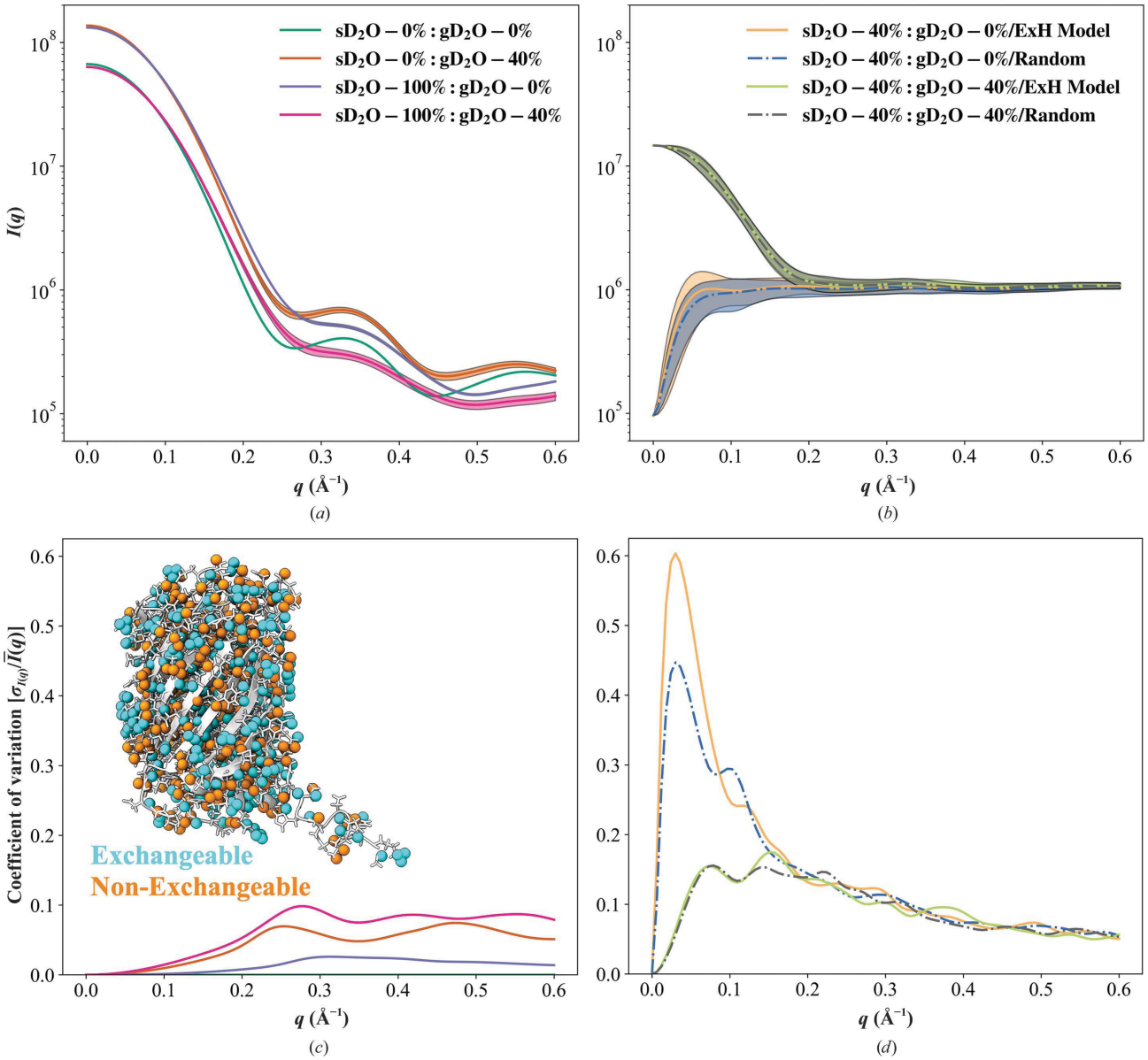


Figure 3
Effect of randomization of the deuteration models on scattering from GFP. (*a*) Comparison of calculated SANS profiles from GFP using 0% and 100% solvent D_2O (sD_2O) with 0% and 40% growth D_2O conditions. The shaded regions are the standard deviation over 100 random deuterations. (*b*) Comparison of the SANS scattering intensity of GFP with solvent randomization for 40% sD_2O and 0% and 40% growth conditions. Shaded regions are the standard deviation. The ExH and Random (40, 40) lines overlap. (*c*) The coefficient of variation (CoV), the ratio of the average scattering intensity to the standard deviation, versus the scattering vector q for the different scattering vectors in (*a*). The structure of GFP is shown with one template of randomizations at 100% sD_2O and 40% gD_2O. Exchangeable H atoms are shown in cyan and non-exchangeable H atoms are colored orange. Two structures showing different deuteration patterns are shown in Supplementary Fig. S2. (*d*) The coefficient of variation (CoV) for scattering profiles in (*b*).

(0% sD_2O, 40% gD_2O) results, indicating that the absolute contrast difference is decreased. Thus, we clearly see a global dependence on the scattering intensities.

At higher q values there is larger variation in the scattering profiles, pointing to a dependence of the scattering on the distribution of contrast in the samples. In the 100% sD_2O and 0% gD_2O (100, 0) scattering there are very small fluctuations from the randomization in the exchangeable hydrogen selection. When non-exchangeable H atoms are added with 40% gD_2O we observe an increase in the CoV, which peaks at 0.28 Å^{-1}, corresponding to an interparticle distance ($2\pi/q$) of 22 Å, the diameter of the β-barrel in GFP. Additionally, there is a shift in the shape of the intensities between (100, 40) and (0, 40) in this region. The peak at 0.33 Å^{-1} in the (0, 40% gD_2O) scattering is depressed relative to the (100, 40) profile, identifying a clear dependence on the explicit deuteration pattern and the contrast. The depression in the shoulder is consistent with a core-shell cylinder model of scattering, with the shell having a higher scattering length density than the core. When comparing *Pepsi-SANS* using bulk deuteration and *SASSENA* (Supplementary Fig. S3) for the (0, 40) and (100, 40) contrasts, we find that *Pepsi-SANS* shows a similar intensity loss at low q compared with *SASSENA* due to bulk contrast but has a different curvature in the profile between 0.3 and 0.4 Å^{-1}. This difference could also be due to the way each program handles the hydration shell: an explicit solvation shell from molecular dynamics in *SASSENA* and grid approximation in *Pepsi-SANS*.

The solvent and the hydration shell can have a large effect on the scattering intensities, making it important to understand solvent deuteration effects. We tested three different schemes for deuteration (Supplementary Fig. S1): an unphysical model with all D atoms assigned to whole D_2O molecules, random deuteration of hydrogens (Random) and a physical model for H–D exchange within the solvent (ExH model) as described in Section 3.2. We tested the solvent deuteration models on a water box with eight NaCl molecules in 0%, 8%, 36% and 100% sD_2O. The 8% and 36% solvent cases (Supplementary Figs. S4*b* and S4*c*) are both special cases in which the scattering length densities of the bulk are zero, *i.e.* '*in vacuo*', and 'transparent' water, where only the oxygen–oxygen structure factor is observed. At 0% and 100% sD_2O (Supplementary Figs. S4*a* and S4*d*) there is no dependence on the model as there is no randomization. In the mixed species the intensities are greater than the homogenous solvent scattering, suggesting there is some excess scattering intensity. Additionally, the excess solvent scattering intensity of the 'D_2O whole' model is larger than both the random and exchange models. We attribute this to the excess average coherent scattering cross section of D_2O being larger than those of HDO and H_2O. As the exchange model restricts a percentage of the total fraction to remain D_2O, we observe a slightly higher average coherent cross section when comparing with pure random exchange. When zooming in to compare the random and exchange models in the insets of Supplementary Figs. S4(*b*) and S4(*c*), there is a difference in the shape of the average scattering profiles and the standard deviations even after 20 randomizations and averaging over 10 ns of simulation time.

When adding the solvent randomization into the calculation of the protein scattering of protiated GFP [Supplementary Figs. S3(*b*) and S3(*d*)], we observe distinct shifts in the intensities and standard deviations in the (40, 0) and (40, 40) scattering. We also compare the random and exchange models to determine whether there is a noticeable difference in the solute scattering. For instance, in the 40% sD_2O and 0% gD_2O [(40, 0)] scattering calculation the forward scattering intensity is three orders of magnitude smaller than the (100, 0) scattering in Fig. 3(*a*), consistent with the well known fact that protiated proteins match out at approximately 42% sD_2O. The solvent is randomized at 40% sD_2O, which increases the total variation in the SANS profiles, as observed in the CoV data (Fig. 3*d*). The low-q peaks (Fig. 3*d*) in the (40, 0) CoV are due to unphysical correlations in the simulation periodic image, while the flatlining at high q in Fig. 3(*c*) is due to excess solvent scattering, as was observed in the solvent-only scattering (Supplementary Fig. S4). This can be removed by running a calculation on a solvent-only box and subtracting the two scattering profiles (data not shown). Interestingly, the CoV decreases when 40% gD_2O is added to GFP, with a maximum at 0.15 Å^{-1}, and decays to the same rough level as the protein-only scattering in Fig. 3(*b*) (top). There is a small difference in the scattering profiles and the CoV of the random versus the exchange models. As such, we expect the physical ExH model of the solvent exchange to provide the most accurate results moving forward, and we implemented the ExH model in the program.

3.2. Accuracy of the prediction of contrast-match points

We have shown above that the explicit deuteration model accurately predicts the fraction of deuteration for non-exchangeable H atoms and that randomization of the deuteration does not affect the $I(0)$ values. We now proceed to test the accuracy of the match-point calculations (Fig. 4). In Fig. 4(*a*), we plot $I(0)^{1/2} \times \mathrm{sign}(\Delta\rho)$ against % sD_2O for lysozyme (PDB entry 1rex; Muraki *et al.*, 1996), maltose-binding protein (PDB entry 1omp; Sharff *et al.*, 1992) and human serum albumin (HSA; SASBDB entry SASDAA6). We compare how our method works with two different SANS calculators: *SASSENA* (solid lines) and *Pepsi-SANS* (dashed lines). The same deuterated PDB templates that were used for *SASSENA* were used for *Pepsi-SANS* with the explicit atom deuteration mode on. In addition to the calculations, we also performed an experimental SANS contrast series for HSA to permit comparison of the contrast-match points determined experimentally with *SCOMAP-XD* (Supplementary Fig. S5).

As expected, the $I(0)^{1/2} \times \mathrm{sign}(\Delta\rho)$ data are linear for maltose-binding protein, lysozyme and HSA as determined from both the SANS calculators. Therefore, we can solve for the zero on the x axis to find the match point. Using *SASSENA* to calculate the scattering profiles, we calculate the match points for lysozyme and HSA at 45% and 41.3% sD_2O, respectively. The match point for MBP produced from 85%

D_2O growth medium is 93% (Dunne *et al.*, 2017). When using *Pepsi-SANS*, the calculated match points are 46%, 42.8% and 96% sD_2O for lysozyme, HSA and MBP, respectively. The experimentally determined match points are shown as dashed vertical lines and are 45% for lysozyme (Stuhrmann & Fuess, 1976), 42.3 ± 0.4% for HSA and 99.5% for MBP (Dunne *et al.*, 2017). *SASSENA* is in excellent agreement for lysozyme and HSA, whereas there is a 6.5% deviation from experiment in the expected match-out condition for MBP. *Pepsi-SANS* performed equally to *SASSENA* for the calculation of match points, indicating that the calculation of match points from 3D structures is generally applicable.

Additional validation of the contrast-match point-fitting procedure was performed with the HSA data calculated from *SASSENA* after performing a scale and offset fit to the experimental data (Supplementary Fig. S5*f*). *SASSENA* obtains excellent agreement with experiment, with a calculated value of 41.6% compared with 42.3%. Performing the fit accounts for factors not considered in the calculation, for example incoherent scattering, enabling us to increase the accuracy of the calculated contrast-match point.

We extended the calculation of match points to several different proteins, nucleic acids and a lipid nanodisc (Supplementary Table S1). Over all of the systems we find that our method produces a root-mean-square error of 2.4% to the experimentally determined match points, with a linear correlation, R^2, value of 0.98 (Fig. 4*b*). Thus, the method and the deuteration models are in excellent agreement with the experimental match points. We tested whether some of the error was due to the H–D exchange model selecting 95% of

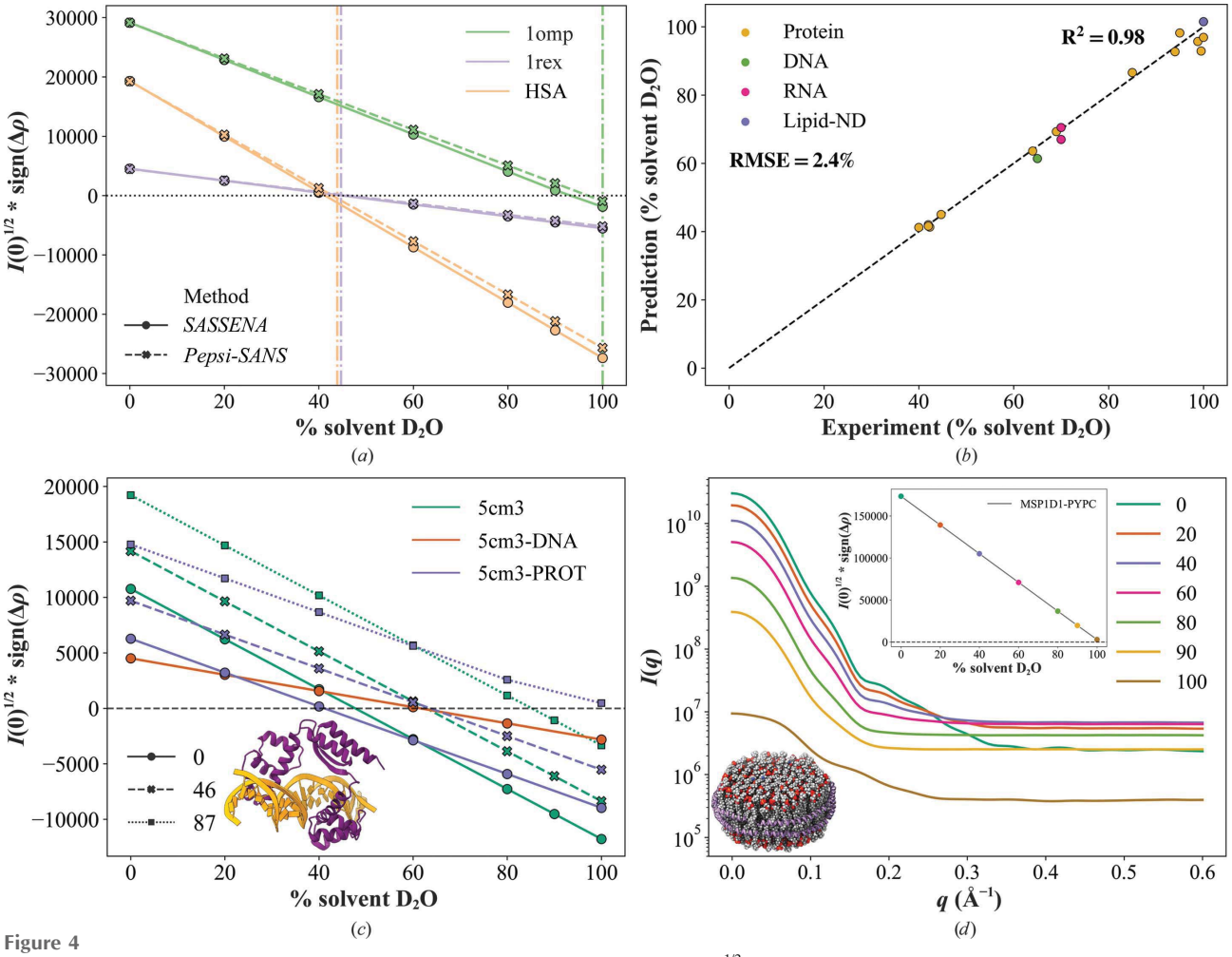


Figure 4
Contrast match-point calculation and SANS profiles. (*a*) Contrast match-point series, $I(0)^{1/2} \times \text{sign}(\Delta\rho)$, versus the solvent D_2O fraction for MBP (PDB entry 1omp; green), HSA (SASBDB entry SASDAA6; orange) and lysozyme (PDB entry 1rex; purple) as well as comparing *SASSENA* (solid circles) and *Pepsi-SANS* (dashed crosses). (*b*) Correlation between experiment and calculated match points for protein (yellow), DNA (green), RNA (magenta) and lipid nanodisc systems (purple). The perfect linear correlation is plotted as a dashed line. The R^2 correlation coefficient is shown in the top right and the root-mean-square error (RMSE) is under the legend. (*c*) Contrast-matching profiles for the KorA–DNA complex (5cm3; teal), DNA only (5cm3-DNA; orange) and protein only (5cm3-PROT; purple). Three protein growth conditions are shown for 0%, 46% and 87% D_2O growth conditions according to the cited deuteration conditions with circles and solid lines, crosses and dashed lines and squares and dotted lines, respectively. (*d*) SANS profiles for the contrast-matching series for the lipid nanodiscs (the structure is shown at the bottom left). The inset for the contrast-match point plot is colored the same as the profiles.

the total number of exchangeable H atoms in the solute (Supplementary Fig. S6). For total exchangeable hydrogen selections of 60% to 100% in increments of 5%, we observe a linear dependence on the match point for PDB entries 1rex and 1oc0 (78% gD_2O; Zhou *et al.*, 2003). The dashed horizontal lines are the experimental match points: 44.7% and 85% sD_2O. Interestingly, both systems would suggest that the exchange factor of 0.95 is too high for PDB entries 1rex and 1oc0, with PDB entry 1rex closer to 0.92 and PDB entry 1oc0 close to 0.85. This error could contribute around a 1–2% difference in the match point. Another source of the RMSE comes from the proteins experimentally determined to match out at 100% sD_2O in 85–87% gD_2O. *SCOMAP-XD* consistently underpredicts the match points at 93% to 97% sD_2O. The exchange-factor error observed above is insignificant in comparison to the difference of 3–7% that would be necessary to match the experimental results. We subsequently performed the inverse contrast-matching calculation by varying the growth conditions from 0% to 100% and solving for the zero in 100% sD_2O (Supplementary Fig. S7). We find that the growth conditions should be 90% D_2O to match out at 100% sD_2O for MBP.

The use of contrast calculators for experimental design is key to successful SANS experiments, particularly in the cases of protein–protein and protein–nucleic acid complexes. We performed a full contrast-match point-series calculation on all of the components for the KorA–DNA binding system from Hyde *et al.* (2017). The scattering profiles for the KorA–DNA (PDB entry 5cm3) complex with 0%, 46% and 87% gD_2O at several sD_2O conditions are shown in Supplementary Fig. S8. We calculated the match-point series for KorA and DNA separately as well as in complex (PDB entry 5cm3; Rajasekar *et al.*, 2016) with KorA expressed in 0%, 46% and 87% gD_2O (70% deuterated), as shown in Fig. 4(*c*). In 0% gD_2O the matched-out solvent D_2O percentages are 47.7%, 61% and 41% for the complex, DNA and protein, respectively. We concurrently see a loss of intensity in the scattering profiles at the match point of the protein (40% sD_2O) and DNA (60% sD_2O) in the complex (PDB entry 5cm3; Supplementary Fig. S8*a*). When 46% and 87% gD_2O are used, the match-out conditions are 63.6% and 96.9% for KorA alone and 62.8% and 85.2% in the complex. Again, we observe the complex match out in the scattering profiles at (60, 46) and the reduced intensity at (100, 87) from the matched-out protein (Supplementary Figs. S8*b* and S8*c*). Thus, under 46% growth conditions we reproduce the conditions which would match out KorA with the DNA within a couple of percent. There was no explicit experimental calculation of the match point of the DNA, but the small error may also be due to the removal of the OrB operon for KorB binding that was included in the experiment.

For the lipid nanodiscs, we performed the deuteration and SANS profile calculations for the 100% sD_2O matched-out MSP1D1–PYPC lipid nanodiscs (Maric *et al.*, 2014) for 0%, 20%, 40%, 60%, 80%, 90% and 100% solvent D_2O (Fig. 4*d*). The calculated match point is 101.5% sD_2O, with a linear dependence on the contrast (inset in Fig. 4*d*). As expected, the SANS profiles reduce in overall intensity as the solvent D_2O percentage approaches the match point, but there is also a clear dependency of other features at distinct q values that diminish. The peak around 0.2 $Å^{-1}$ is maximal when the contrast is the highest, 0% sD_2O, but slowly diminishes and is completely gone by 60% sD_2O. The corresponding distance correlation to the peak is between the head groups of the opposite bilayers, so by changing the deuteration of the head groups relative to the solvent and the lipid acyl chains this distance could be highlighted in a SANS experiment.

The use of selective deuteration, targeting specific amino acids or domains with deuteration to ascertain distances between components within a complex using different contrasts, is another key tool for characterizing biomolecular complexes. We built three different constructs of the TIA1 RNA-binding protein (Sonntag *et al.*, 2017): the protiated construct ($R1_H$:$R2R3_H$), the R1 deuterated construct ($R1_D$:$R2R3_H$) and the R2R3 deuterated system ($R1_H$:$R2R3_D$). We illustrate the TIA1 R1 and R2R3 domains in Supplementary Fig. S9(*a*). We performed the match-point calculation and compared it with the experimental results (dashed vertical lines) in Supplementary Fig. S9(*b*). In $R1_H$:$R2R3_H$ and $R1_D$:$R2R3_H$ there is an almost perfect agreement with the experimentally determined match points, while that for the $R1_H$:$R2R3_D$ construct is 1.3% lower than the experimental result of 94% at 92.7%.

3.3. Identifying *q*-dependent contrast effects using Debye decomposition

In previous sections, we identified specific differences in the shape of the scattering profiles upon the addition of deuterium through H–D exchange. For a deeper understanding of the structural features and the domains associated with these domains, we calculated the scattering contribution of each component and the cross terms of the KorA–DNA complex, using the Debye formula, for two contrast conditions: (60, 0) (Fig. 5*a*) and (100, 87) (Fig. 5*b*). As mentioned earlier, the (60, 0) contrast is designed to match out the DNA in the complex, while the (100, 87) contrast will match out KorA. The Debye formula allows one to calculate the scattering profiles from a set of pairwise distances. In this case, we calculate the pairwise distances between protein–protein, protein–DNA and DNA–DNA atoms. Comparing the Debye calculation for the sum of all the contributions (Debye-Full; orange) with the *SCOMAP-XD* calculation (blue), we observe that the scattering profiles are similar, with features appearing at 0.2 $Å^{-1}$ (31.4 Å). The protein–protein distances contribute the most to this feature as the Debye-Full and protein–protein profiles are similar. As expected, the DNA–DNA contribution is flat as it is matched out, but the protein–DNA distances do contribute between 0.2 and 0.3 $Å^{-1}$, although at a much lower intensity than the protein–protein distances. Conversely, in the (100, 87) contrast (Fig. 5*b*) the dominant contribution is the DNA–DNA distances (purple). Interestingly, the protein–protein distances still contribute a small intensity at 0.2 $Å^{-1}$, while the cross terms only contribute at low q and 0.3 $Å^{-1}$

(20.9 Å) not at 0.2 $Å^{-1}$ as in the (60, 0) contrast. With different contrasts, we clearly observe q-dependent changes in the scattering profiles from the different macromolecular components.

4. Discussion and conclusions

Contrast-variation methods for SANS through changing the D_2O concentration in the solvent or the incorporation of D into the target solute through H–D exchange, growth of cells in D_2O or specific deuteration has repeatedly been shown to be a powerful tool for understanding the solution structures of proteins, lipids, nucleic acids and their respective complexes. The use of contrast calculators to design SANS experiments is a vital addition to the neutron scatterer's repertoire for efficient and accurate experiments. Popular predictors such as *SASSIE* (Sarachan *et al.*, 2013) and *MULCh* (Whitten *et al.*, 2008) have been very accurate in predicting bulk match-out conditions, but rely on the average total fraction of deuteration due to H–D exchange and non-exchangeable hydrogen properties as well as in the solvent. There has, however, been little investigation of q-dependent contrast effects, which go beyond the bulk-scattering properties and which require spatially resolved and structural understanding of deuteration (Ibel & Stuhrmann, 1975; Witz, 1983). In this study, we have designed a workflow and explicit deuteration models for both solute and solvent to calculate bulk contrast matching of biomolecular systems. Our code predicted the contrast-match points of several systems within 2.4% sD_2O and demonstrated that it can reproduce the experimental design of a protein–DNA system. Importantly, we showed how using the same deuteration models developed in *SCOMAP-XD* with the Debye formula on the same protein–DNA system identifies distinct differences in the high-q scattering profiles, suggesting q-dependent contrast effects.

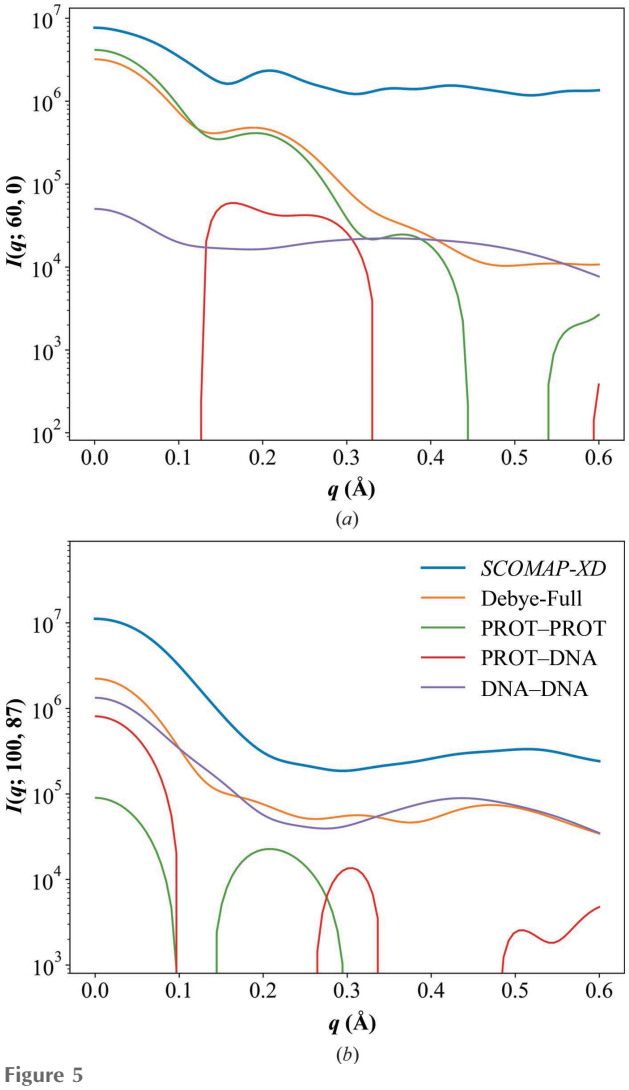


Figure 5
Debye decomposition of solute components for PDB entry 5cm3 for (60% sD_2O, 0% gD_2O) (*a*) and (100% sD_2O, 87% gD_2O) (*b*). The scattering curves calculated using *SCOMAP-XD* are shown in blue. The sum of the scattering contributions (Debye-Full) from the protein–protein (PROT–PROT; green), protein–DNA (PROT–DNA; red) and DNA–DNA (purple) pairwise distances is displayed in orange.

The computational deuteration procedure developed here considers all three methods by which deuterium can be incorporated into biomacromolecular systems: using a structure-based H–D exchange model, random non-exchangeable hydrogen selection from an empirical model built from mass-spectrometry data and a direct selection methodology. The H–D exchange model is well established and has recently been extended to understand the incorporation of deuterium into proteins over time (Pedersen *et al.*, 2019). The H–D exchange thermodynamic model worked well for the understanding of the contrast-match points of proteins, nucleic acids and their complexes in this study. For natural isotopic abundance biomolecules, H–D exchange accounts for the contrast-match points of nucleic acids and proteins at 65–70% solvent D_2O and 39–45% solvent D_2O, respectively. Our method accurately predicts such contrast-match points compared with experimentally determined contrast-match points within a couple of percent (Supplementary Table S1).

We assumed that an existing model for backbone amide H–D exchange also extends to the side chains of amino acids and nucleic acids. The conventional assumption in SANS contrast matching is that the side chains always exchange proportionally to the % sD_2O, while the backbone exchanges at 95% of the % sD_2O on average. Our extended structural model for H–D exchange to include side chains seems to have little effect on the calculated match points since the contrast-match point does not depend on the placement of the deuterium, only the average amount (see Fig. 3*b*). As no structural model currently exists to understand the incorporation of deuterium into non-exchangeable hydrogen sites during biosynthesis, we used a random selection of all available C—H sites in the solute. We first used mass spectrometry to determine the exact fraction of deuterium incorporated into GFP for several different growth conditions. Our results reproduce prior experiments showing there is a quadratic dependence in the incorporation of deuterium into proteins,

which has been attributed to the kinetic isotope effect (Lederer *et al.*, 1986). We fitted our mass-spectroscopy results to a quadratic function and found that the final fits give c_1 and c_2 parameters in equation (4) of 0.40 and 0.46, respectively, which are close to the parameters, 0.42 and 0.44, found for the DNA-dependent RNA polymerase from *E. coli* (Lederer *et al.*, 1986). Our model calculates the fraction of deuteration at a 100% growth condition to be 86.86%, which is in excellent agreement with previous work on peptide deformylase (Leiting *et al.*, 1998). We expect these types of models to form a foundation for further understanding of deuterium incorporation into biomacromolecules and its interpretation in SANS.

In both H–D exchange and growth-condition models, we observed no dependence on the forward scattering and the contrast-match points from varying the deuteration patterns in the solute, consistent with theory. However, there was a clear dependence in the SANS profiles, particularly in the mid-q to high-q regions. There were both bulk effects, *i.e.* shifts in the overall intensity of the scattering profiles, as well as local effects of deuteration. Local contrast effects were observed when comparing *SASSENA* and *Pepsi-SANS*, where the scattering dampened the peak at q = 0.33 Å^{-1} when going from 0% solvent D_2O to 100% D_2O with 40% growth D_2O (Fig. 3*a*). Local effects in the SANS profiles were captured in randomization models, with the coefficient of variation (Figs. 3*b* and 3*d*) indicating that the q values occur at intra-protein length scales. In GFP, the CoV peaked at scattering vectors associated with the diameter of the β-barrel. This may be due to the higher localization of deuterium at the surface-exposed exchangeable H atoms versus the interior of the protein resulting in a shift in the local contrast relative to the solvent. The largest limitation in observing these features is from the incoherent scattering of the solvent. The incoherent scattering can dominate the mid-q to high-q scattering regions, thereby washing out potential key features. More accurate models of deuteration, better neutron scattering instrumentation and contrast-matching techniques may be necessary to resolve this limitation.

The 2.5% difference between the experimental and theoretically calculated match points could be due to several factors. The fraction of H–D exchange is typically understood to comprise around 90% of the backbone amide H atoms and all of the side-chain exchangeable H atoms (Perkins, 1986). However, recent work has suggested that depending on the time of incubation the backbone amide fraction could be as low as 80% (Pedersen *et al.*, 2019). In terms of the direct calculation of match points performed in this work, the total fraction of exchangeable H atoms has a small effect on the contrast, changing around 0.5% solvent D_2O for a 5% change in the total fraction of exchange (Supplementary Fig. S6). In turn, the random deuteration of non-exchangeable H atoms using the empirical model from mass spectrometry gave excellent results, but fundamentally there could be an amino-acid residue-specific deuteration pattern. Previous work on *E. coli* DNA-dependent RNA polymerase (Lederer *et al.*, 1986) identified some trends in the deuteration patterns, but a preliminary analysis based on residue properties in this study showed no particular trend in the deuteration patterns from mass spectrometry. Neutron protein crystallography has shown that aromatic residues are less likely to be deuterated, but further studies need to be performed to deconvolve the amino-acid-specific dependence from their metabolic pathways. In this study, we only considered glycerol as a carbon source in the culture medium. The addition of deuterated or nondeuterated carbon sources to the culture medium will affect the total deuteration percentage (Leiting *et al.*, 1998; Weiss *et al.*, 2021). This would be another source of the 2.5% error, and potentially the errors at higher gD_2O conditions. In general, there is also error in the experimental determination for the match points propagating from errors in the molecular weights, Guinier fits, incomplete buffer subtraction or potential aggregation in samples due to high D_2O concentrations (Jeffries *et al.*, 2016; Rubinson *et al.*, 2008). Published experimental errors for contrast-match points are typically within 1–5%, so our results here are well within experimental error. In the case of nucleic acids, experimental match points of 65% for DNA and 70% for RNA are typically used, from which our predictions have some of the largest deviations in this study. Overall, we demonstrate that our method gives excellent agreement for bulk contrast-match point calculations.

Structure-based SANS calculators, both explicit and implicit solvent, have been a huge aid for the understanding of neutron scattering and biomolecular complexes together. One of the main limitations of our method is the necessity for a structure. However, with the advent of *AlphaFold*2 and high-resolution cryo-EM and X-ray crystallographic data, the availability of template structures to determine starting contrast results is common. As we showed, the structure itself has limited if no effect on the bulk contrast-match points. Further techniques such as MD simulations or SAXS can be used to identify distances of interest for q-dependent contrast matching for binding studies of small molecules or complexes. The implementation of a highly parallel explicit solvent SANS calculator in *SASSENA* allows very large systems to be studied if a structure is available. Fundamentally, we believe that the structure-based methods developed here can easily be expanded to a larger variety of biomacromolecular systems as well as polymers.

5. Related literature

The following references are cited in the supporting information for this article: Abel *et al.* (1996), Braun *et al.* (2011), Falb *et al.* (2010), Hennig *et al.* (2013, 2014), Okuda *et al.* (2021) and Puster *et al.* (2019).

Acknowledgements

AH would like to thank M. D. Smith for help with *SASSENA*. This research used resources at the High Flux Isotope Reactor and Spallation Neutron Source, a DOE Office of Science User Facility operated by Oak Ridge National Laboratory. This research used resources of the Compute and Data Environ-

ment for Science (CADES) at Oak Ridge National Laboratory, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC05-00OR22725.

Funding information

This work is supported by project ERKPA14 funded by the Office of Biological and Environmental Research (BER) in the Department of Energy (DOE) Office of Science. Neutron scattering experiments on Bio-SANS were supported by the Center for Structural Molecular Biology funded by DOE BER project ERKP291.

References

Abel, K., Yoder, M. D., Hilgenfeld, R. & Jurnak, F. (1996). *Structure*, **4**, 1153–1159.
Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015). *SoftwareX*, **1–2**, 19–25.
Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. (1993). *Proteins*, **17**, 75–86.
Best, R. B. & Vendruscolo, M. (2006). *Structure*, **14**, 97–106.
Braun, N., Zacharias, M., Peschek, J., Kastenmüller, A., Zou, J., Hanzlik, M., Haslbeck, M., Rappsilber, J., Buchner, J. & Weinkauf, S. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 20491–20496.
Chen, P.-C. & Hub, J. S. (2014). *Biophys. J.* **107**, 435–447.
Chen, P.-C., Shevchuk, R., Strnad, F. M., Lorenz, C., Karge, L., Gilles, R., Stadler, A. M., Hennig, J. & Hub, J. S. (2019). *J. Chem. Theory Comput.* **15**, 4687–4698.
Darden, T., York, D. & Pedersen, L. (1993). *J. Chem. Phys.* **98**, 10089–10092.
Dunne, O., Weidenhaupt, M., Callow, P., Martel, A., Moulin, M., Perkins, S. J., Haertlein, M. & Forsyth, V. T. (2017). *Eur. Biophys. J.* **46**, 425–432.
Engelman, D. M. & Moore, P. B. (1975). *Annu. Rev. Biophys. Bioeng.* **4**, 219–241.
Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995). *J. Chem. Phys.* **103**, 8577–8593.
Falb, M., Amata, I., Gabel, F., Simon, B. & Carlomagno, T. (2010). *Nucleic Acids Res.* **38**, 6274–6285.
Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.
Grudinin, S., Garkavenko, M. & Kazennov, A. (2017). *Acta Cryst.* D**73**, 449–464.
Heller, W. T., Hetrick, J., Bilheux, J., Calvo, J. M. B., Chen, W.-R., DeBeer-Schmitt, L., Do, C., Doucet, M., Fitzsimmons, M. R., Godoy, W. F., Granroth, G. E., Hahn, S., He, L., Islam, F., Lin, J., Littrell, K. C., McDonnell, M., McGaha, J., Peterson, P. F., Pingali, S. V., Qian, S., Savici, A. T., Shang, Y., Stanley, C. B., Urban, V. S., Whitfield, R. E., Zhang, C., Zhou, W., Billings, J. J., Cuneo, M. J., Leal, R. M. F., Wang, T. & Wu, B. (2022). *SoftwareX*, **19**, 101101.
Heller, W. T., Urban, V. S., Lynn, G. W., Weiss, K. L., O'Neill, H. M., Pingali, S. V., Qian, S., Littrell, K. C., Melnichenko, Y. B., Buchanan, M. V., Selby, D. L., Wignall, G. D., Butler, P. D. & Myles, D. A. (2014). *J. Appl. Cryst.* **47**, 1238–1246.
Hennig, J., Militti, C., Popowicz, G. M., Wang, I., Sonntag, M., Geerlof, A., Gabel, F., Gebauer, F. & Sattler, M. (2014). *Nature*, **515**, 287–290.
Hennig, J., Wang, I., Sonntag, M., Gabel, F. & Sattler, M. (2013). *J. Biomol. NMR*, **56**, 17–30.
Hess, B. (2008). *J. Chem. Theory Comput.* **4**, 116–122.
Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. (1997). *J. Comput. Chem.* **18**, 1463–1472.
Hoover, W. G. (1985). *Phys. Rev. A*, **31**, 1695–1697.
Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H. & MacKerell, A. D. (2017). *Nat. Methods*, **14**, 71–73.
Hyde, E. I., Callow, P., Rajasekar, K. V., Timmins, P., Patel, T. R., Siligardi, G., Hussain, R., White, S. A., Thomas, C. M. & Scott, D. J. (2017). *Biochem. J.* **474**, 3121–3135.
Ibel, K. & Stuhrmann, H. B. (1975). *J. Mol. Biol.* **93**, 255–265.
Jacrot, B. (1976). *Rep. Prog. Phys.* **39**, 911–953.
Jeffries, C. M., Graewert, M. A., Blanchet, C. E., Langley, D. B., Whitten, A. E. & Svergun, D. I. (2016). *Nat. Protoc.* **11**, 2122–2153.
Jo, S., Kim, T., Iyer, V. G. & Im, W. (2008). *J. Comput. Chem.* **29**, 1859–1865.
Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *J. Chem. Phys.* **79**, 926–935.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.
Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. (2020). *Protein Sci.* **29**, 66–75.
Kim, H. S. & Gabel, F. (2015). *Acta Cryst.* D**71**, 57–66.
Krueger, S. (2022). *Curr. Opin. Struct. Biol.* **74**, 102375.
Laux, V., Callow, P., Svergun, D. I., Timmins, P. A., Forsyth, V. T. & Haertlein, M. (2008). *Eur. Biophys. J.* **37**, 815–822.
Lederer, H., May, R. P., Kjems, J. K., Schaefer, W., Crespi, H. L. & Heumann, H. (1986). *Eur. J. Biochem.* **156**, 655–659.
Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, C. L., MacKerell, A. D., Klauda, J. B. & Im, W. (2016). *J. Chem. Theory Comput.* **12**, 405–413.
Leiting, B., Marsilio, F. & O'Connell, J. F. (1998). *Anal. Biochem.* **265**, 351–355.
Lindner, B. & Smith, J. C. (2012). *Comput. Phys. Commun.* **183**, 1491–1501.
Maric, S., Skar-Gislinge, N., Midtgaard, S., Thygesen, M. B., Schiller, J., Frielinghaus, H., Moulin, M., Haertlein, M., Forsyth, V. T., Pomorski, T. G. & Arleth, L. (2014). *Acta Cryst.* D**70**, 317–328.
Martyna, G. J., Klein, M. L. & Tuckerman, M. (1992). *J. Chem. Phys.* **97**, 2635–2643.
Max, J.-J. & Chapados, C. (2002). *J. Chem. Phys.* **116**, 4626–4642.
McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J. & Pande, V. S. (2015). *Biophys. J.* **109**, 1528–1532.
Merzel, F. & Smith, J. C. (2002*a*). *Acta Cryst.* D**58**, 242–249.
Merzel, F. & Smith, J. C. (2002*b*). *Proc. Natl Acad. Sci. USA*, **99**, 5378–5383.
Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. (2022). *Nat. Methods*, **19**, 679–682.
Moore, P. B. (1977). *Anal. Biochem.* **82**, 101–108.
Muraki, M., Harata, K., Sugita, N. & Sato, K. (1996). *Biochemistry*, **35**, 13562–13567.
Nosé, S. (1984). *Mol. Phys.* **52**, 255–268.
Okuda, A., Inoue, R., Morishima, K., Saio, T., Yunoki, Y., Yagi-Utsumi, M., Yagi, H., Shimizu, M., Sato, N., Urade, R., Kato, K. & Sugiyama, M. (2021). *Biophys. Physicobiol.* **18**, 16–27.
Park, S., Bardhan, J. P., Roux, B. & Makowski, L. (2009). *J. Chem. Phys.* **130**, 134114.
Parrinello, M. & Rahman, A. (1981). *J. Appl. Phys.* **52**, 7182–7190.
Pedersen, M. C., Wang, Y., Tidemand, F. G., Martel, A., Lindorff-Larsen, K. & Arleth, L. (2019). *J. Appl. Cryst.* **52**, 1427–1436.
Perkins, S. J. (1981). *Biochem. J.* **199**, 163–170.
Perkins, S. J. (1986). *Eur. J. Biochem.* **157**, 169–180.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H. & Ferrin, T. E. (2021). *Protein Sci.* **30**, 70–82.
Powles, J. G., Dore, J. C. & Page, D. I. (1972). *Mol. Phys.* **24**, 1025–1037.
Puster, L. O., Stanley, C. B., Uversky, V. N., Curtis, J. E., Krueger, S., Chu, Y. & Peterson, C. B. (2019). *Biochemistry*, **58**, 5117–5134.
Qi, Y., Lee, J., Klauda, J. B. & Im, W. (2019). *J. Comput. Chem.* **40**, 893–899.
Rajasekar, K. V., Lovering, A. L., Dancea, F., Scott, D. J., Harris, S. A., Bingle, L. E., Roessle, M., Thomas, C. M., Hyde, E. I. & White, S. A. (2016). *Nucleic Acids Res.* **44**, 4947–4956.
Rubinson, K. A., Stanley, C. & Krueger, S. (2008). *J. Appl. Cryst.* **41**, 456–465.
Sarachan, K. L., Curtis, J. E. & Krueger, S. (2013). *J. Appl. Cryst.* **46**, 1889–1893.
Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quiocho, F. A. (1992). *Biochemistry*, **31**, 10657–10663.
Sonntag, M., Jagtap, P. K. A., Simon, B., Appavou, M. S., Geerlof, A., Stehle, R., Gabel, F., Hennig, J. & Sattler, M. (2017). *Angew. Chem. Int. Ed.* **56**, 9322–9325.
Stuhrmann, H. B. & Fuess, H. (1976). *Acta Cryst.* A**32**, 67–74.
Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.
Timmins, P. A. & Zaccai, G. (1988). *Eur. Biophys. J.* **15**, 257–268.
Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. C. (2005). *J. Comput. Chem.* **26**, 1701–1718.
Wang, I., Hennig, J., Jagtap, P. K., Sonntag, M., Valcárcel, J. & Sattler, M. (2014). *Nucleic Acids Res.* **42**, 5949–5966.
Weiss, K. L., Fan, Y., Abraham, P., Odom, M., Pant, S., Zhang, Q. & O'Neill, H. (2021). *Methods Enzymol.* **659**, 219–240.
Whitten, A. E., Cai, S. & Trewhella, J. (2008). *J. Appl. Cryst.* **41**, 222–226.
Whitten, A. E., Jacques, D. A., Hammouda, B., Hanley, T., King, G. F., Guss, J. M., Trewhella, J. & Langley, D. B. (2007). *J. Mol. Biol.* **368**, 407–420.
Witz, J. (1983). *Acta Cryst.* A**39**, 706–711.
Yang, F., Moss, L. G. & Phillips, G. N. (1996). *Nat. Biotechnol.* **14**, 1246–1251.
Zhang, Q., Willison, L. N., Tripathi, P., Sathe, S. K., Roux, K. H., Emmett, M. R., Blakney, G. T., Zhang, H.-M. & Marshall, A. G. (2011). *Anal. Chem.* **83**, 7129–7136.
Zhou, A., Huntington, J. A., Pannu, N. S., Carrell, R. W. & Read, R. J. (2003). *Nat. Struct. Mol. Biol.* **10**, 541–544.