JCTC Journal of Chemical Theory and Computation

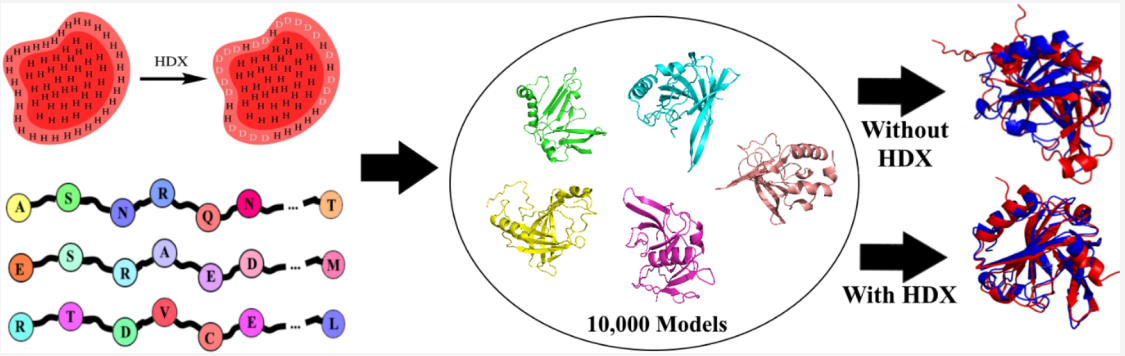
pubs.acs.org/JCTC Article

Protein Structure Prediction from NMR Hydrogen–Deuterium Exchange Data

Daniel R. Marzolf, Justin T. Seffernick, and Steffen Lindert*

Cite This: J. Chem. Theory Comput. 2021, 17, 2619–2629 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

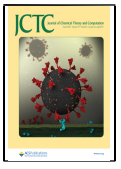


ABSTRACT: Amide hydrogen–deuterium exchange (HDX) has long been used to determine regional flexibility and binding sites in proteins; however, the data are too sparse for full structural characterization. Experiments that measure HDX rates, such as HDX-NMR, have far higher throughput compared to structure determination via X-ray crystallography, cryo-EM, or a full suite of NMR experiments. Data from HDX-NMR experiments encode information on the protein structure, making HDX a prime candidate to be supplemented by computational algorithms for protein structure prediction. We have developed a methodology to incorporate HDX-NMR data into *ab initio* protein structure prediction using the Rosetta software framework to predict structures based on experimental agreement. To demonstrate the efficacy of our algorithm, we examined 38 proteins with HDX-NMR data available, comparing the predicted model with and without the incorporation of HDX data into scoring. The root-mean-square deviation (rmsd, a measure of the average atomic distance between superimposed models) of the predicted model improved by 1.42 Å on average after incorporating the HDX-NMR data into scoring. The average rmsd improvement for the proteins where the selected model rmsd changed after incorporating HDX data was 3.63 Å, including one improvement of more than 11 Å and seven proteins improving by greater than 4 Å, with 12/15 proteins improving overall. Additionally, for independent verification, two proteins that were not part of the original benchmark were scored including HDX data, with a dramatic improvement of the selected model rmsd of nearly 9 Å for one of the proteins. Moreover, we have developed a confidence metric allowing us to successfully identify near-native models in the absence of a native structure. Improvement in model selection with a strong confidence measure demonstrates that protein structure prediction with HDX-NMR is a powerful tool which can be performed with minimal additional computational strain and expense.

INTRODUCTION

The function of a protein is dictated by its structure; thus, the understanding of biological processes is significantly facilitated by the knowledge of the protein structure.[1] Despite this, the gap between the number of known sequences of proteins and their three-dimensional structures is widening by the day.[2] There are many experimental approaches available for structural and dynamic characterization of proteins. In the world of dynamic studies, one such method is the monitoring of hydrogen–deuterium exchange rates via nuclear magnetic resonance spectroscopy (HDX-NMR). HDX-NMR data are typically generated to elucidate regional flexibility or binding sites after a protein's structure has been fully characterized via other methods, such as X-ray crystallography, cryo-EM, or a full suite of NMR structural experiments.[3−6] The result of HDX-NMR experiments is a residue-resolved map of exchange rates, allowing for extrapolation of regional flexibility and solvent exposure, two factors that are generally considered to influence the HDX rate.[7] HDX rate determination is not exclusive to NMR, however, and can be measured using mass spectrometry (MS) as well.[8,9] While HDX-NMR studies yield

Received: January 21, 2021
Published: March 29, 2021



https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619–2629

important information on structure and dynamics, these are still sparse data, generally insufficient for full protein structure determination or unambiguous dynamics characterization. Computational methods that can facilitate the structural interpretation of HDX-NMR data are required.

There have been strides to incorporate experimental techniques with computation, with efforts spanning back to the 1980s with NMR and X-ray crystallography and more recently EPR, MS, and cryo-EM among others.[10−30] HDX experiments, originally probed in the 1970s,[31] have been used to map exchange rates onto atomic-resolution structures to assign dynamic properties to otherwise static representations.[4,32−37] In the general case, HDX rates have also been coupled to molecular dynamics simulations to explain variation in different regions of a protein.[14,38−42] Additionally, these data have been incorporated into protein−protein docking of complexes with known tertiary structure to elucidate quaternary structure.[43−45] However, importantly, HDX rates have not yet been used to predict *de novo* tertiary structure. Previous implementations for structural characterization rely on either homology modeling or some starting structures such as an alternative conformation of a protein or a designed protein.[46−49] While there are multiple software packages with impressive results that exist for *ab initio* structure prediction, such as the co-evolution-dependent neural network AlphaFold,[50] the secondary structure assembling BCL,[51] or iterative threading I-TASSER,[52] none have been coupled to experimental data as frequently or diversely as the Rosetta Modeling Software.[13,20,21,27,30,53−62] Rosetta *ab initio* structure prediction allows for the generation of models from amino acid sequence alone, assembling fragments generated from short segments with similar sequences using Monte Carlo sampling combined with a hybrid classical physics and probabilistic knowledge-based scoring function in both coarse-grained and full-atom modeling, similar to other multiscale modeling methods.[63,64] Due to its modular score function, Rosetta is an ideal candidate to use HDX-NMR data for *ab initio* structure prediction.

In this work, we have developed methods to account for residual solvent exposure, through amide neighbor count (NC) and residual relative solvent accessible surface area, and flexibility, through hydrogen-bonding energies and order score (OS), all within the Rosetta framework. While the Rosetta *ab initio* sampling approach does not allow for the determination of realistic folding assembly pathways, the ability to quantify the flexibility and exposure of residues in a native-state model means that correlations between HDX of residues within the protein and the protein structure can be explored. Using HDX-NMR data for 38 proteins from the Start2Fold database,[65] we have developed a score term for the Rosetta energy function based upon agreement with experimental data, while also accounting for local sequence context. Using this new scoring term, we have scored structures generated using Rosetta's *ab initio* prediction application, improving the root-mean-square deviation (RSMD, a measure of average atomic distance between superimposed models) from native of the best scoring predicted model with negligible additional computational expense; in several predictions, rmsd improved by more than 5 Å, including one prediction improving by over 11 Å.

MATERIALS AND METHODS

Benchmark Dataset. We assembled a benchmark dataset of proteins with HDX-NMR data from the Start2Fold database, a curated database for experimental HDX-NMR determinations of folding pathways and regional stability.[65] The available data were provided in the form of per-residue classification for stability experiments. The experimental stability was classified as either strong for residues that were highly protected from exchange, weak for residues which exchanged quickly, or medium for residues in between ranges. Each category was defined by the database in accordance with the measured experimental data, such as protection factor (a measure inversely proportional to the exchange rate constant) or change in peak intensity over time. For example, a strong residue has a higher protection factor compared to a medium or weak residue. Due to more strict restrictions on the class, only data for the strong residues were used in our analysis, as these were often residues which did not exchange at all, whereas a weak residue could transiently move and be exchanged, which would not be relevant to the static model. Of the 57 proteins available (at the time of search) in the Start2Fold database, 38 were chosen for the scoring benchmark because they contained residues classified into the strong category, were monomeric in solution, and had an experimentally determined structure in the Protein Data Bank (PDB), and models with less than 10 Å rmsd from native were sampled with Rosetta (protocol described below). Separately, two proteins were selected from the Start2Fold database to serve as an independent verification set, to test the scoring protocol outside of the benchmark set. Protein lengths ranged from 56 to 179. A summary of the benchmark set is shown in Table S1.

Model Generation. For each of the 38 proteins in the benchmark set, 10,000 decoy models were generated using Rosetta's standard AbinitioRelax protocol.[66−71] For this, files containing 3-mer and 9-mer residual fragments were generated using the Robetta Web server.[72] These fragments were used as an input in a Monte Carlo assembly, where structures were scored using coarse-grain energy functions, followed by all-atom relaxation and use of the Ref2015 scoring function in the final full-atom refinement.[63] Cα-rmsd from native was calculated for each of the generated decoys for use in benchmarking. The rmsd of only ordered secondary structured elements (SSEs) was calculated using a custom PyMOL script which aligned a truncated PDB of the native structure containing only the ordered SSEs to the generated models.[73] The number of 10,000 decoy models was chosen because the AbinitioRelax protocol generally requires the generation of a large number of models to adequately sample conformational space. The structure with the lowest (most favorable) Rosetta score was identified as the predicted structure.

Calculation of Flexibility and Exposure Metrics. There is a general consensus in the HDX community that HDX rates are dependent on the local flexibility and solvent exposure at the amide hydrogen position.[7] Thus, we have calculated parameters which quantify these features on a per-residue basis for use in the scoring of Rosetta decoy models based on HDX-NMR data. All calculations were performed using Rosetta applications. Four parameters were chosen to quantitatively represent flexibility (hydrogen-bond energy and OS) and exposure (NC and relative solvent accessible surface area).

To quantify flexibility, the backbone hydrogen-bond energy (H-Bond) was extracted from the Rosetta Energy Function via the residue_energy_breakdown application, using the following hydrogen-bond energy terms: short-range (in sequence) backbone−backbone interactions (hbond_sr_bb), long-range

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

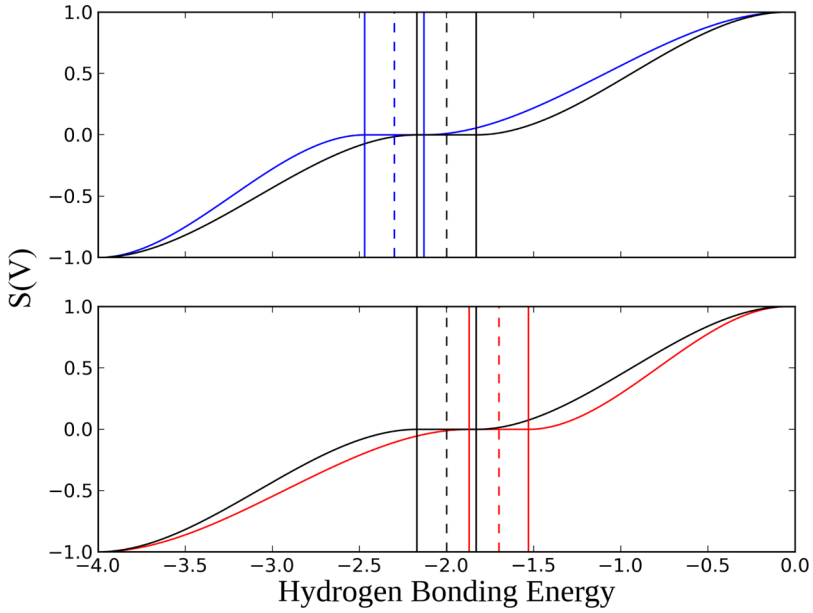


Figure 1. (Top) Score function $S(V)$ for the H-bond parameter with (blue) and without (black) an HDX-catalyzing SAP factor incorporated into score range definition. (Bottom) Score function $S(V)$ for the H-bond parameter with (red) and without (black) an HDX-inhibiting SAP factor incorporated into the score range definition. Solid vertical lines mark the borders of the nonzero scoring range for their respective colors. Dotted lines indicate the mean (black) and mean + SAP factor (blue/red).

(in sequence) backbone−backbone interactions (hbond_lr_bb), and backbone-side chain interactions (hbond_bb_sc).[63] The term involving hydrogen-bonding interactions between side chains (hbond_sc) was discarded for calculations performed herein due to the transient nature of these interactions caused by side chain flexibility and for the lack of involvement of the amide proton in these interactions. The hydrogen-bonding energies were extracted such that the only energy contribution was from the backbone amide group rather than a sum of backbone amide and carbonyl oxygen contributions. To do so, a Rosetta application (ragul_find_all_hbonds) was used to determine the donor/acceptor pair for each residue. Using the determined pairing, the energy was extracted only if the amide hydrogen was involved in the interaction, rather than the carbonyl oxygen. The energies of the terms were summed to generate the final residual hydrogen-bonding energy, though typically, only one energy term was nonzero. The expected trend of this hydrogen-bonding energy is that a lower HDX rate would correlate with a higher magnitude of the energy (i.e. more negative). As a second measure of flexibility, OS (a measure of residue-resolved disorder) was calculated using the Rosetta Residue-Disorder application that calculates a window-averaged Rosetta score to map per-residue disorder.[74,75] We used the Ref2015 scoring function and a window size of 11 to quantify disorder of residue i, based on a score average spanning from residue $i-5$ to residue $i+5$. Similar to hydrogen bonding, as the HDX rate decreases, the OS is expected to become more negative.

With respect to exposure, relative per-residue solvent accessible surface area (RelSASA) was calculated using Rosetta scoring classes (SasaCalc). Since RelSASA decreases as exposure decreases, the expected correlation between the HDX rate and RelSASA is that as the HDX rate decreases, the RelSASA is expected to trend toward zero. For NC calculations, the Rosetta NC application (per_residue_solvent_exposure) was modified in order to calculate conical NC based on the oxygen atoms neighboring the amide proton, as the oxygen atoms can both sterically and electronically alter the amide proton environment.[16] The angle cutoffs were chosen such that no atoms behind the amide were counted as neighbors, with the angle contribution midpoint set at $\pi/2$ radians. The distance contribution midpoint was set to 9 Å, as optimized previously.[24] NC increases as exposure decreases; thus, as the HDX rate decreases, the NC should increase, as the amide becomes less accessible to deuterated solvents.

HDX Score. For each of the four calculated parameters, the mean and standard deviations for each strength category for the 38 native crystal structures were determined to verify expected trends between the parameters and protection strength categories. In order to score decoys based on agreement with HDX data, a scoring function for each calculated metric was developed to reward or penalize residues of *ab initio* models. This was done by scoring residues in the strong category based on the deviation of calculated metrics from the distribution observed in the crystal structures. The strategy was to reward residues that strongly matched hypothesized features of strong residues and penalize residues that did not.

If a calculated metric was within a range around the average, as defined in eq 1 (where μ is the mean of the native distribution of the parameter, σ is the standard deviation of the native distribution of the parameter, and f is a scaling factor of the standard deviation), the residue was scored as zero. Figure 1 (solid black vertical lines) shows an example (for H-bond energy) of the range where a zero score was applied (average shown as a dotted black vertical line).

(The value of f changed depending on the range of the native structure distribution. e.g., the NC distribution ranged from ∼2 to 18, and the standard deviation of the distribution was 2.25; due to the size of the range compared to the standard deviation, the value of f was set to 1.0 for NC, resulting in a zero-score region of size 4.5, or approximately 28% of the total range. Conversely, RelSASA's distribution ranged from 0 to 1, with a standard deviation of 0.26; if the f value was set to 1.0

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

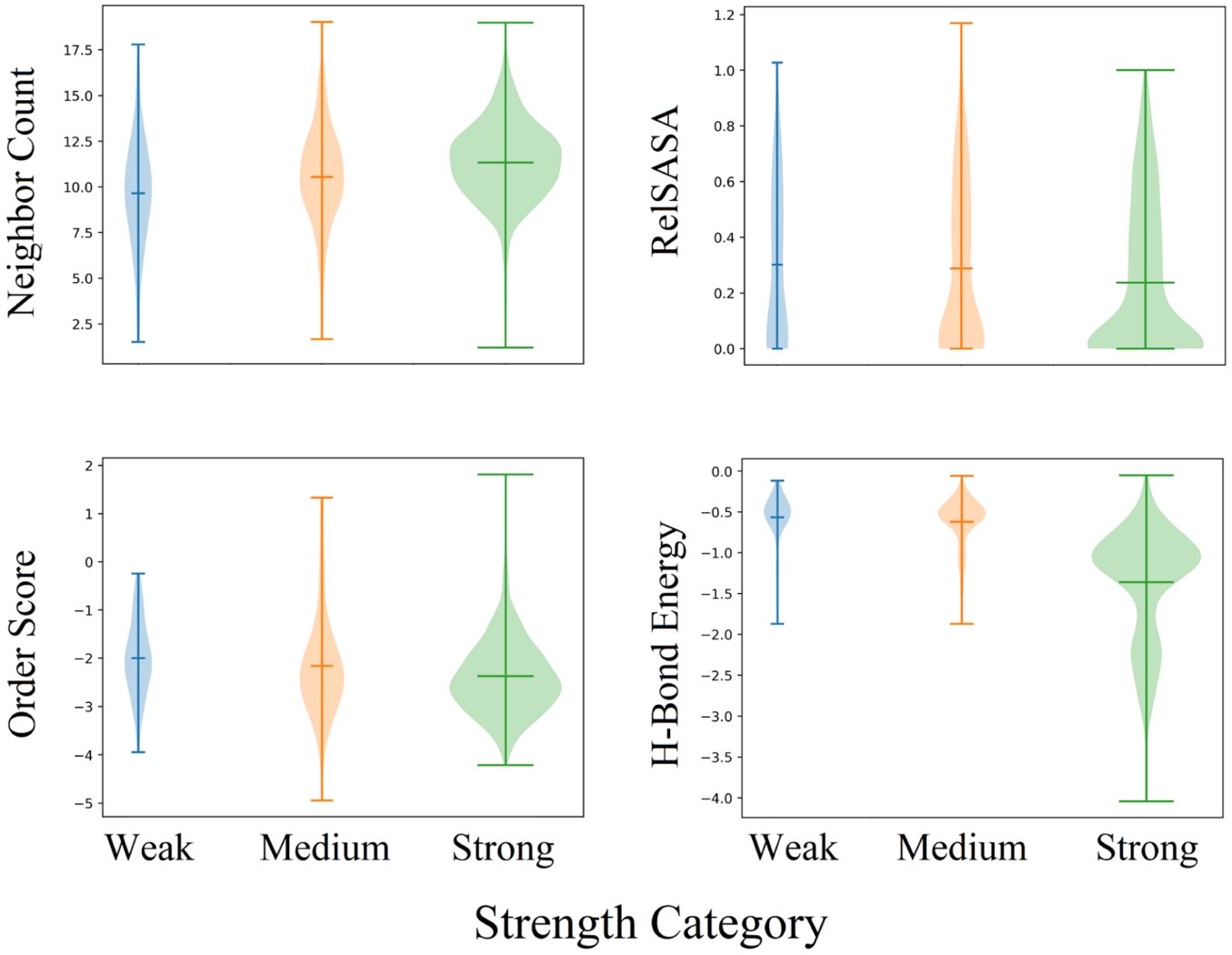


Figure 2. Distributions of calculated parameters for native structures with widths proportional to the dataset size. Horizontal line in the center of each distribution marks the mean of the dataset.

for RelSASA, the zero-score region would include all values in a range from 0 to 0.52, which contains approximately 80% of strong residues. To prevent this, the *f* value was set to 0.25 for RelSASA, resulting in only 13% of the distribution being a zero-score and allowing for residues with very low RelSASA to be scored. For similar reasons, the *f* values for H-bond energies and OS were set to 0.25 and 0.5, respectively.)

$$\text{Range} = \mu \pm \sigma^{*}f \tag{1}$$

However, outside of this range, residues of a decoy model were rewarded or penalized based on the level of (dis)-agreement with the distribution of the calculated parameters of the native structure and our structural hypotheses. (The specific function will be described in more detail at the end of this section.) For example, following the hypothesis that a residue within the strong category should be less flexible, if a strong residue in a generated model had a high H-bond energy (outside of the zero-score region), it resulted in a penalty for that model which increases as the H-bond energy increases. This penalty function is shown in Figure 1 (at high H-bond energies). Conversely, a strong residue with a low (more negative) H-bond energy (outside of the zero-score region) would be rewarded. Figure 1 (at low H-bond energies) shows the function used for the reward.

HDX has been shown experimentally to be sequence dependent as well, where the level of exchange depends on the identity of side chains adjacent to the amide in sequence.[76] Thus, another feature, coined "side chain amide protection (SAP)", was used to influence scoring, altering the scoring range based on sequence context. To generate the SAP factor, the relative literature exchange rates in an acid catalyzed environment, which were derived from the residue identity, were added together for the i and $i - 1$ residues, as the side chains for the those residues affect the HDX rate.[76] For example, the 28th residue of 1BDD, a strong residue, is an arginine with a SAP value of −0.59, and the 27th residue is glutamine with a SAP value of −0.27; thus, the SAP factor for R28 in 1BDD was −0.86. This sequence-specific score range adaptation allowed us to account for the intrinsic HDX protection or catalysis from a local sequence when scoring individual residues. All residual SAP values are provided in a separate supplemental file. With the implementation of the SAP factor, the scoring range was defined by eq 2, where SAP is the residual SAP value, P is the SAP scaling factor (P = 0.2 for RelSASA, 3.0 for H-bond, and not applied to OS and NC), and other terms follow the naming conventions *vide supra*.

$$\text{Range} = \mu \pm \sigma^{*}f + \text{SAP}^{*}P \tag{2}$$

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

The scoring function outside of the zero-score region was implemented as a set of two fade functions as qualitatively described previously (function shown in eq 3), graphically depicted in Figure 1 for H-bond with examples of HDX-catalyzing SAP (top) and HDX-inhibiting SAP (bottom, where V is the calculated parameter value (NC, OS, RelSASA, or H-bond energies), C is the nearest range cutoff value ($\mu \pm \sigma^* f$ + SAP), and M is the nearest extreme value of the distribution. If the SAP factor indicated that neighboring side chains were HDX-catalyzing groups, the nonzero scoring cutoffs would shift such that the penalty region expanded (Figure 1, top, vertical blue lines), while HDX-inhibiting groups would expand the reward region (Figure 1, bottom, vertical red lines).

$$S(V) = 2^*\left(-\frac{V - C - (M - C)}{M - C}\right)^3 - 3 \\ ^*\left(-\frac{V - C - (M - C)}{M - C}\right)^2 + 1 \quad (3)$$

The HDX score was defined as a weighted sum of the Rosetta score and the score components for solvent accessibility (NC, RelSASA) and flexibility (OS, H-bond) as shown in eq 4, where RS is the Rosetta score, $S(V)$ is the score derived from each category, and all other variables follow naming conventions from above. The results of this scoring were relatively stable with respect to different combinations of weights.

$$\text{HDX score} = \text{RS} + 1^*S(\text{NC}) + 5^*S(\text{RelSASA}) + 5 \\ ^*S(\text{OS}) + 30^*S(\text{Hbond}) \quad (4)$$

P_{near} (a measurement of how funnel-like a score vs rmsd distribution is) was calculated for each of the distributions using the Rosetta score without HDX data incorporated and using the HDX score (eq 4).[77] P_{near} values can range from 0 (indicating a poor funnel with several low-energy models at a range of RSMDs from native) to 1 (a perfect funnel with a unique low-energy conformation in the near-native state). For the calculation of P_{near}, λ was set equal to 2.0 and K_BT was set equal to 1.0.

Confidence Metric. To determine confidence in model selection, a confidence metric was developed based upon knowledge independent of the native structure. The confidence metric was defined as the average rmsd of the top 100 scoring models to the top scoring model. This was chosen because a low average rmsd indicated high structural similarity for the top scoring models. We hypothesized that this would suggest a favorable energy landscape and thus better scoring of native-like structures. Therefore, if the confidence metric was less than 5 Å, predictions were identified as high confidence, and if the metric was above 5 Å, predictions were identified as low confidence.

RESULTS AND DISCUSSION

Experimental Data from Native Structures Follow Hypothesized Exposure and Flexibility Trends. Given the consensus in the HDX community of the influence of both exposure and flexibility on HDX rates, the parameters to quantify these residual properties were calculated from the native crystal structures for each of the 38 benchmark proteins to determine if the relationship between strength categories matched our hypotheses.[7] Figure 2 depicts the distributions of all calculated parameters (NC, RelSASA, OS, and H-bond) as a function of residual HDX protection. Each of the averages of the distributions followed the hypothesized trends, where the strong category corresponded to the lowest exposure and flexibility. The averages for the strong category for the parameters were 11.41 (NC), 0.23 (RelSASA), −2.39 (OS), and −1.36 (H-bond).

Scoring using HDX was performed solely using residues within the strong category for three reasons. First, the size of the dataset for the strong category was significantly larger than other strength categories, with 678 residues in the strong category compared to 165 and 267 residues in the weak and medium categories, respectively. Additionally, while a weak residue could be in a highly dynamic region where the residue could change from exposed to buried via random motion, a strong residue must be resistant to exchange for the majority of the experiment, resulting in a more reliable metric to generate HDX restraints for modeling. Finally, we observed minimal overlap between the strong category and the weaker categories, especially in the extremes of the distribution, which were used in the scoring algorithm; this provided a higher confidence that the distribution which the scoring is based upon is unique to the strong category rather than one where a value of a parameter could be weak or strong.

Initial Rosetta Model Generation Yielded a Large Distribution of High and Low RMSD Models. For each of the 38 proteins selected from the Start2Fold database,[65] 10,000 models were generated using Rosetta's standard AbinitioRelax protocol.[66−70] While near-native structures (rmsd < 3 Å) were predicted using the Rosetta score without HDX data incorporated for 11 proteins, the average rmsd of the predicted structure was 6.68 Å (Table S1). The rmsd of the predicted structure was greater than 5 Å for 18 benchmark proteins and greater than 10 Å for 8 proteins. However, for 32 of the 38 proteins, at least one model with rmsd less than 5 Å was sampled with Rosetta *ab initio*, and for 22/38, at least one model with rmsd less than 3 Å was sampled. This indicated that near-native structure selection was possible for a majority of the benchmark set if an additional score was used. P_{near} (a measurement of how funnel-like a score vs rmsd distribution is) values were generally low, indicating that models of high and low rmsd had similar energies, with an average P_{near} of 0.136.

Individual HDX Parameter-Based Scoring Improved Model Selection Accuracy. We developed the HDX scoring function as a linear combination between the Rosetta Ref2015 scoring function and our newly developed terms that quantify the agreement with HDX data based on exposure and flexibility parameters. If, in a generated model, the exposure or flexibility parameters of residue agreed with the distribution of the parameters in the X-ray crystal structures, and the residue was rewarded, with those opposite penalized. The score was dependent upon the level of (dis)agreement to the distribution. Additionally, the nonzero scoring range was modulated by the SAP factor, which accounted for sequence context by biasing scoring toward reward or penalty depending on the side chains immediately neighboring the amide proton. Before incorporation into the linear combination, each of the individual parameters was used to score based on HDX rate agreement. In doing so, each of the parameters was analyzed to determine whether scores based on the hypothesized trends could be used to improve model selection alone, as well as give insights into which, if any, of the parameters were the most

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

beneficial. Moreover, the parameters needed to be tested to determine which would benefit from the inclusion of the SAP factor.

Results of each scoring method are listed in Table 1. Scoring using a static scoring range cutoff based solely upon the mean and standard deviation of the distribution of the parameters in the native structures (without the SAP factor) were somewhat unimpressive. We hypothesized this to be due to the lack of sequence context, where an amide neighboring two glycine side chains would be treated the same as one surrounded by a phenylalanine and tyrosine, neglecting the differences in the steric and electronic environment between side chains. Thus, the SAP factor was introduced to create a sequence dependent scoring range that would account for a residue with minimal neighbors being sterically or electronically hindered from HDX due to its neighboring side chains in sequence rather than the full environment measured by the parameters.

Table 1. Summary of Results following Scoring for the Neighbor Count (NC)-, Relative Solvent Accessible Surface Area (RelSASA)-, Order Score (OS)-, and Hydrogen Bond Energy (H-Bond)-Based HDX Scoring, with and without the Inclusion of the SAP Factor in the Definition of the Score Range (eqs 1 and 2)[a]

	without SAP				
	weight applied to Rosetta score	average Δrmsd of all proteins (Å)	average Δrmsd of proteins with changes in rmsd (Å)	number of proteins with rmsd improved	number of proteins with rmsd increased
NC	10	−0.73	−2.90	10	0
RelSASA	5	−0.74	−1.41	8	5
OS	6	−0.22	−1.92	4	0
H-bond	10	−0.02	−0.11	6	3
	with SAP				
	weight applied to Rosetta score	average Δrmsd of all proteins (Å)	average Δrmsd of proteins with changes in rmsd (Å)	number of proteins with rmsd improved	number of proteins with rmsd increased
NC	10	−0.63	−2.22	9	2
RelSASA	5	−0.93	−1.61	9	4
OS	6	−0.20	−1.39	5	0
H-bond	10	−0.94	−4.00	9	0

[a]A change in rmsd is defined as the magnitude of selected model rmsd change by greater than 0.5 Å.

Each parameter was tested with (eq 2) and without (eq 1) including sequence context to modify the scoring range cutoffs, as shown in Table 1. Model selection was improved when SAP was included in the RelSASA and H-bond-based scoring. Conversely, NC and OS methods did not benefit from including the SAP factor.

The ineffectiveness of the SAP factor for NC- and OS-based scoring was initially surprising and contrary to our hypothesis that a sequence dependent scoring range would improve model selection. However, this effect can be explained by elements that determine NC and OS compared to H-bond and RelSASA. NC is dependent on the location of oxygen atoms within a hemisphere surrounding the amide proton, with the contribution to the NC degrading with respect to the distance and angle to the amide NH vector. Similarly, OS is calculated by a window-averaged Rosetta per-residue score, dependent on the five residues on the N- and C-terminal sides of the scored amide and each of their local environments. However, SAP is based solely on the i and $i-1$ side chains, far closer in both space and sequence than the NC and OS determinants. Thus, the inclusion of SAP to these terms did not benefit scoring. Conversely, H-bond and RelSASA are inherently dependent on the local environment in the location of i and $i-1$ residues. This is supported by the improvement in the RelSASA- and H-bond-based scoring when the SAP factor was included, as these parameters are determined by the scored residue alone and the amide proton's H-bond partner, respectively. Thus, for all HDX scoring methods, the SAP factor was excluded from OS- and NC-based scoring methods, while it was included in H-bond- and RelSASA-based scoring, as indicated in Table 1. Further comparison of the results of SAP inclusion can be found in the Supporting Information.

Figure 3 shows the results of scoring using the individual parameter-based score compared to using the Rosetta score without HDX data incorporated. In general, model selection improved for each of the scoring methods, with an average improvement of 0.71 Å. Importantly, while some parameter-based scoring resulted in an increase in selected model rmsd, this deficiency was not shared by other parameters. For example, if the rmsd of the selected model was higher when the NC scoring term was used, the selected model had the same or better rmsd when scored based on another parameter. The individual parameters were categorized based on whether they quantified flexibility (OS and H-bond) or solvent exposure (NC or RelSASA). The scoring results of the combination of these terms are discussed in the Supporting Information, Figure S1, and Table S2. In general, these paired terms performed slightly better than the terms separately.

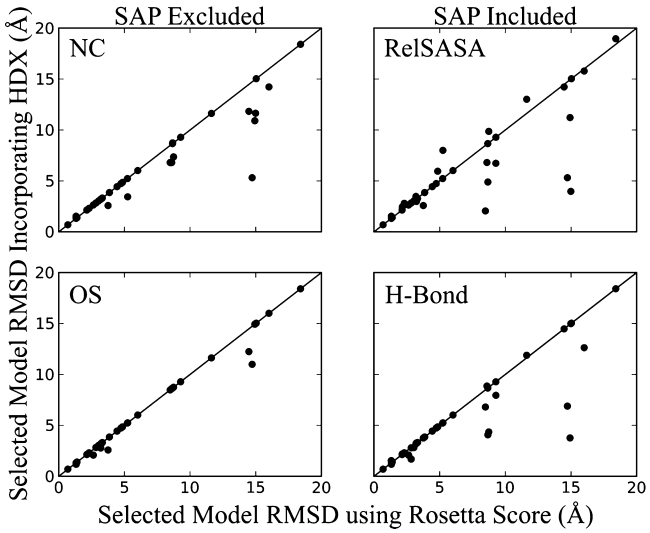


Figure 3. Selected model rmsd for the Rosetta prediction and scoring that incorporated HDX data, using NC (top, left), RelSASA (relative solvent accessible surface area) (top, right), OS (order score) (bottom, left), or H-Bond (amide hydrogen bonding energy) (bottom, right) as the parameters used to determine agreement. The SAP factor was included in only the RelSASA and H-bond scoring (right). Markers below the $y = x$ line indicate a protein with an improvement in the selected model rmsd with those above indicating the worsening.

Combination of All Four Score Terms Produced the Largest Improvement in Model Selection. The final HDX

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

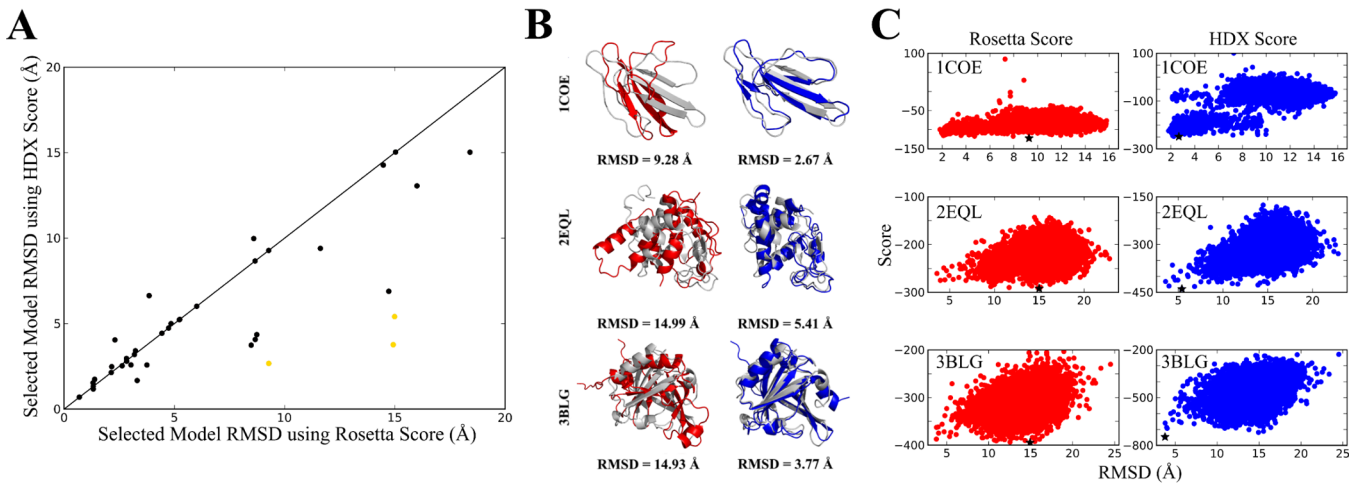


Figure 4. Results of the 38 protein benchmark set using the HDX score. (A) Selected model rmsd for the Rosetta prediction and when scoring with the HDX score. Markers below the $y = x$ line indicate a protein with an improvement in selected model rmsd with those above indicating the worsening. Points in gold are represented in (B,C). (B) Lowest scoring models (red) using the Rosetta (left) and HDX (right) scores overlaid with the X-ray crystal structure (blue). (C) Rosetta (left) and HDX (right) score vs rmsd plots of three proteins that benefited by use of the HDX score. The lowest scoring model is marked by a black star.

score (eq 4) was composed of a weighted sum of the individual terms that measure exposure and flexibility. The weighting of the terms skewed highly toward the hydrogen bonding component, one of the measurements of residual flexibility. This is to be expected due to the mechanism of HDX; if the amide proton is engaged in an energetically favorable hydrogen bond, it is less likely to undergo reactions requiring electron transfer. Thus, the presence of a highly stabilizing hydrogen bond is known to correlate strongly to the exchange rate and thus the experimental HDX category.[78]

Figure 4A shows the selected model rmsd when using the Rosetta score compared to the HDX score. The average improvement of the rmsd of the best scoring model was 1.42 Å, with seven proteins improving by over 4 Å. The selected models for four of these proteins using the Rosetta score and the HDX score are overlaid with the native structure in Figure 4B. When using the HDX score, for proteins with greater than 0.5 Å rmsd difference between selected model with and without HDX data, 12/15 improved, with an average improvement of 3.63 Å. For two proteins, the top scoring model using the HDX score was the best possible model from the decoy pool (lowest rmsd), including one case where the rmsd of the predicted structure improved from 14.93 to 3.77 Å (Table S1) when HDX data was included. Additionally, while the overall selected model rmsd improved by 1.42 Å, the average rmsd of residues within ordered secondary structure elements improved by 0.92 Å, with a maximal improvement of 10.48 Å, indicating that the improvement in rmsd was not solely in disordered regions. These regions are important to protein function yet highly dynamic compared to core regions which are less likely to have major disruptions in solution and are vital to the protein structure as well.[79−81] However, ideally, incorporating the HDX data would improve the model selection for every protein, such a result would require far less sparse experimental data, removing the benefit of pairing high-throughput computation and experimentation. However, this sparse HDX NMR dataset was able to improve prediction in cases when the score distribution from the initial prediction without experimental data was close to accurate.

Not only did the top scoring model improve when HDX data were included, the average rmsd of the top 10 scoring models also improved from 6.97 Å using the Rosetta score to 6.30 Å, shown in Figure S2 for all proteins. While only one of the average rmsds increased by greater than 0.5 Å, the average rmsd of the top 10 scoring models improved by more than 0.5 Å for sixteen proteins, indicating strong model selection improvement. Figure S3 shows the rmsd distribution for the top 10 scoring models for all proteins in the dataset. When using the HDX score, the rmsd distribution shifted toward a lower rmsd compared to using the Rosetta score without HDX data incorporated, with a marked improvement in the number of models in the sub-5 Å range. Figure 4C shows the score versus rmsd distributions for three proteins for which we observed significant improvement in rmsd of the top 10 scoring models upon application of the HDX score. Among all distributions in the benchmark set, P_{near} improved by 7% when the HDX score was used compared to when the Rosetta score was used, another indication that model selection improvement was not limited to only the top scoring model.

While the native structures and thus rmsds were known for the models generated within the benchmark dataset, this knowledge is unavailable for true *ab initio* prediction, motivating the establishment of a confidence metric which can be used as a marker of a probable near-native model generation. To this end, we developed a confidence metric, the average rmsd of the top 100 scoring models to the top scoring model when using the HDX score. Figure 5 shows the selected model rmsd as a function of our confidence measure. If the average rmsd to the top scoring model was less than 5.0 Å (indicating strong funneling and thus high confidence), the average selected model rmsd was 2.54 Å. The rmsd of the selected model for all 18 proteins identified by the metric as high confidence was less than 5 Å. Additionally, all proteins with a selected model rmsd below 2.5 Å were identified as high confidence. Contrasting this, the average rmsd of proteins in the low confidence region was 7.70 Å, with 14 of the 20 proteins selecting a model with an rmsd of 5 Å or above. The distinct difference in model selection quality between the high and low confidence regions indicates that the confidence

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

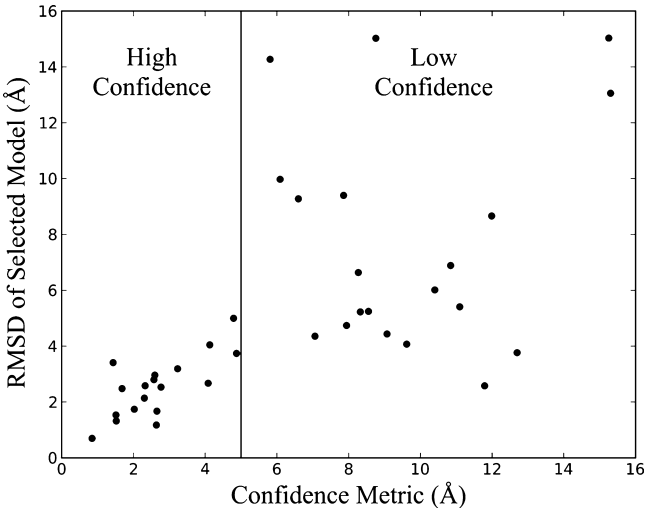


Figure 5. Plot of the confidence metric (the average rmsd of the top 100 scoring models to the selected model using the HDX score) vs rmsd of the selected model following scoring using the HDX score, where the solid line indicates the confidence cutoff of 5.0 Å such that proteins with an average rmsd to the left of the line have a high confidence for increased model selection accuracy.

measure is a powerful tool for enabling positive identification of near-native models, even in the absence of a known native structure.

HDX Score Improved Model Selection for Proteins Outside of the Benchmark Set. To ensure broader applicability, two proteins were selected from the Start2Fold database for independent verification separately from the benchmark set. These proteins (PDB IDs: 1A2P and 1HRC) matched the requirements of the benchmark set (monomeric in solution, had an experimentally determined structure in the PDB, and models with less than 10 Å rmsd from native were sampled with Rosetta). The HDX score was calculated for the proteins as stated above. When using the Rosetta score, the selected model rmsd from native for 1A2P was 12.87 Å and for 1HRC was 13.34 Å. However, using the HDX score, 1A2P remained approximately the same, selecting a model with an rmsd from native of 13.09 Å, while 1HRC improved to 4.41 Å, selecting the best model generated in the pool. The selected models for 1HRC using the Rosetta score and HDX score overlaid with the X-ray crystal structure are shown in Figure 6. Score versus rmsd distributions are shown in Figure S4.

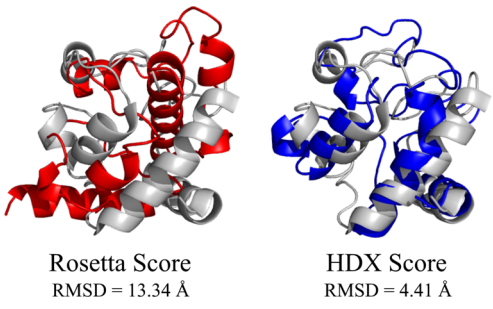


Figure 6. Selected models (red) using the Rosetta score (left) and HDX score (right) overlaid with the X-ray crystal structure (blue).

CONCLUSIONS

HDX rates have been studied for decades, primarily to characterize the dynamics of proteins which had already been structurally elucidated experimentally. Though *ab initio* protein structure prediction has made major strides in a similar timeframe, moving from computing small peptides to deep-learning structural prediction, this too often requires a broadly inaccessible amount of computational power or conjunction with expensive and difficult experimentation.[2,82,83] We sought to eliminate this burden by utilizing data from high-throughput, broadly accessible HDX experiments that are too sparse for structure determination themselves, but, as we have demonstrated, highly useful when incorporated into computational analysis and structural prediction.

To our knowledge, we are the first to incorporate sparse HDX-NMR data into computational *ab initio* protein structure prediction. By incorporating HDX data into Rosetta scoring, the rmsd of the selected model improved by 1.42 Å on average; of the 15 proteins whose rmsd changed by greater than 0.5 Å, 12 improved with an average improvement of 3.63 Å. The rmsd of the selected model also improved for core residues in ordered secondary structure elements by 0.92 Å, with an improvement as high as 10.48 Å. Additionally, a confidence metric was developed to determine the confidence of identifying native-like predicted structure. The rmsd of the selected model for all 18 proteins in the high-confidence region was less than 5 Å. Improvement in model selection with a strong confidence measure demonstrates that protein structure prediction with HDX-NMR is a powerful tool in facilitating protein structure determination.

While HDX-MS has recently gained popularity as a method of HDX rate determination, the large dataset available via the Start2Fold database made HDX-NMR ideal for the development of a scoring system. Importantly, the scoring algorithm developed from this database paves the way for expansion to HDX-MS data as well as multimeric structural prediction. While HDX-MS typically generates fragment-resolved (as opposed to residue-resolved) data, the HDX principles are maintained regardless of the experiment, making HDX-MS a prime target for adaptation of the scoring algorithm. Mass spectrometry is typically far higher throughput than NMR experimentation, which would increase the overall speed of this prediction method. Moreover, MS experiments are not as stringently bound to protein size limitations as NMR experiments, which tend to be unviable for proteins larger than 50 kDa unless specialized sampling is used, which has its own set of limitations.[84] Removing the size limitation allows for studies of complex structures via differential HDX-MS experiments. Future work may focus on expanding our scoring algorithm to HDX-MS for monomeric structure prediction and protein complex structure prediction which are crucial to the vast majority of biological processes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00077.

Performance of individual and paired components of the final score term; a summary of the proteins in the benchmark set; and results for the top 10 selected models using the HDX score and the independent verification test (PDF)

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

Details of residue number, residue type, and SAP value (TXT)

AUTHOR INFORMATION

Corresponding Author

Steffen Lindert − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-3976-3473; Phone: 614-292-8284; Email: lindert.1@osu.edu; Fax: 614-292-1685

Authors

Daniel R. Marzolf − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States*

Justin T. Seffernick − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00077

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank the members of the Lindert group for useful discussions. The authors would like to thank the Ohio Supercomputer Center for valuable computational resources.[85] Research reported in this publication was supported by NSF (CHE 1750666) to S.L.

REFERENCES

(1) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P., *Molecular Biology of the Cell*, 4th ed.; Garland Science: New York, 2002.

(2) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.

(3) Krishna, M.; Hoang, L.; Lin, Y.; Englander, S. W. Hydrogen exchange methods to study protein folding. *Methods* **2004**, *34*, 51−64.

(4) Uzawa, T.; Nishimura, C.; Akiyama, S.; Ishimori, K.; Takahashi, S.; Dyson, H. J.; Wright, P. E. Hierarchical folding mechanism of apomyoglobin revealed by ultra-fast H/D exchange coupled with 2D NMR. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13859−13864.

(5) Pérez, J. M. J.; Renisio, J. G.; Prompers, J. J.; van Platerink, C. J.; Cambillau, C.; Darbon, H.; Frenken, L. G. J. Thermal Unfolding of a Llama Antibody Fragment: A Two-State Reversible Process†. *Biochemistry* **2001**, *40*, 74−83.

(6) Xu, S.; Ni, S.; Kennedy, M. A. NMR Analysis of Amide Hydrogen Exchange Rates in a Pentapeptide-Repeat Protein from A. thaliana. *Biophys. J.* **2017**, *112*, 2075−2088.

(7) Vadas, O.; Burke, J. E. Probing the dynamic regulation of peripheral membrane proteins using hydrogen deuterium exchange-MS (HDX-MS). *Biochem. Soc. Trans.* **2015**, *43*, 773−786.

(8) Bai, Y. Protein Folding Pathways Studied by Pulsed- and Native-State Hydrogen Exchange. *Chem. Rev.* **2006**, *106*, 1757−1768.

(9) Masson, G. R.; Burke, J. E.; Ahn, N. G.; Anand, G. S.; Borchers, C.; Brier, S.; Bou-Assaf, G. M.; Engen, J. R.; Englander, S. W.; Faber, J.; Garlish, R.; Griffin, P. R.; Gross, M. L.; Guttman, M.; Hamuro, Y.; Heck, A. J. R.; Houde, D.; Iacob, R. E.; Jørgensen, T. J. D.; Kaltashov, I. A.; Klinman, J. P.; Konermann, L.; Man, P.; Mayne, L.; Pascal, B. D.; Reichmann, D.; Skehel, M.; Snijder, J.; Strutzenberg, T. S.; Underbakke, E. S.; Wagner, C.; Wales, T. E.; Walters, B. T.; Weis, D. D.; Wilson, D. J.; Wintrode, P. L.; Zhang, Z.; Zheng, J.; Schriemer, D. C.; Rand, K. D. Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **2019**, *16*, 595−602.

(10) Srivastava, A.; Nagai, T.; Srivastava, A.; Miyashita, O.; Tama, F. Role of computational methods in going beyond X-ray crystallography to explore protein structure and dynamics. *Int. J. Mol. Sci.* **2018**, *19*, 3401.

(11) Pilla, K. B.; Gaalswyk, K.; MacCallum, J. L. Molecular modeling of biomolecules by paramagnetic NMR and computational hybrid methods. *Biochim. Biophys. Acta Protein Proteonomics* **2017**, *1865*, 1654−1663.

(12) Ardenkjaer-Larsen, J.-H.; Boebinger, G. S.; Comment, A.; Duckett, S.; Edison, A. S.; Engelke, F.; Griesinger, C.; Griffin, R. G.; Hilty, C.; Maeda, H.; Parigi, G.; Prisner, T.; Ravera, E.; van Bentum, J.; Vega, S.; Webb, A.; Luchinat, C.; Schwalbe, H.; Frydman, L. Facing and Overcoming Sensitivity Challenges in Biomolecular NMR Spectroscopy. *Angew. Chem. Int. Ed.* **2015**, *54*, 9162−9185.

(13) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685.

(14) Konermann, L.; Pan, J.; Liu, Y.-H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **2011**, *40*, 1224−1234.

(15) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure* **2010**, *18*, 923−933.

(16) Aprahamian, M. L.; Lindert, S. Utility of Covalent Labeling Mass Spectrometry Data in Protein Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2019**, *15*, 3410−3424.

(17) Harvey, S. R.; Seffernick, J. T.; Quintyn, R. S.; Song, Y.; Ju, Y.; Yan, J.; Sahasrabuddhe, A. N.; Norris, A.; Zhou, M.; Behrman, E. J.; Lindert, S.; Wysocki, V. H. Relative interfacial cleavage energetics of protein complexes revealed by surface collisions. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 8143−8148.

(18) Leelananda, S. P.; Lindert, S. Iterative Molecular Dynamics-Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. *J. Chem. Theory Comput.* **2017**, *13*, 5131−5145.

(19) Lindert, S.; Hofmann, T.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. Ab initio protein modeling into CryoEM density maps using EM-Fold. *Biopolymers* **2012**, *97*, 669−677.

(20) DiMaio, F.; Tyka, M. D.; Baker, M. L.; Chiu, W.; Baker, D. Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* **2009**, *392*, 181−190.

(21) Kahraman, A.; Herzog, F.; Leitner, A.; Rosenberger, G.; Aebersold, R.; Malmström, L. Cross-link guided molecular modeling with ROSETTA. *PloS One* **2013**, *8*, No. e73411.

(22) Hauri, S.; Khakzad, H.; Happonen, L.; Teleman, J.; Malmström, J.; Malmström, L. Rapid determination of quaternary protein structures in complex biological samples. *Nat. Commun.* **2019**, *10*, 192.

(23) Whitley, J. A.; Ex-Willey, A. M.; Marzolf, D. R.; Ackermann, M. A.; Tongen, A. L.; Kokhan, O.; Wright, N. T. Obscurin is a semi-flexible molecule in solution. *Protein Sci.* **2019**, *28*, 717−726.

(24) Aprahamian, M. L.; Chea, E. E.; Jones, L. M.; Lindert, S. Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data. *Anal. Chem.* **2018**, *90*, 7721−7729.

(25) Webb, B.; Viswanath, S.; Bonomi, M.; Pellarin, R.; Greenberg, C. H.; Saltzberg, D.; Sali, A. Integrative structure modeling with the Integrative Modeling Platform. *Protein Sci.* **2018**, *27*, 245−258.

(26) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **2012**, *10*, No. e1001244.

(27) Leelananda, S. P.; Lindert, S. Using NMR Chemical Shifts and Cryo-EM Density Restraints in Iterative Rosetta-MD Protein Structure Refinement. *J. Chem. Inf. Model.* **2020**, *60*, 2522−2532.
(28) Biehn, S. E.; Lindert, S. Accurate protein structure prediction with hydroxyl radical protein footprinting data. *Nat. Commun.* **2021**, *12*, 341.
(29) Seffernick, J. T.; Lindert, S. Hybrid methods for combined experimental and computational determination of protein structure. *J. Chem. Phys.* **2020**, *153*, 240901.
(30) Seffernick, J. T.; Harvey, S. R.; Wysocki, V. H.; Lindert, S. Predicting Protein Complex Structure from Surface-Induced Dissociation Mass Spectrometry Data. *ACS Cent. Sci.* **2019**, *5*, 1330−1341.
(31) Rosa, J. J.; Richards, F. M. An experimental procedure for increasing the structural resolution of chemical hydrogen-exchange measurements on proteins: Application to ribonuclease S peptide. *J. Mol. Biol.* **1979**, *133*, 399−416.
(32) Palmer, A. G. Probing molecular motion by NMR. *Curr. Opin. Struct. Biol.* **1997**, *7*, 732−737.
(33) Hooke, S. D.; Radford, S. E.; Dobson, C. M. The Refolding of Human Lysozyme: A Comparison with the Structurally Homologous Hen Lysozyme. *Biochemistry* **1994**, *33*, 5867−5876.
(34) Choe, S. E.; Matsudaira, P. T.; Osterhout, J.; Wagner, G.; Shakhnovich, E. I. Folding Kinetics of Villin 14T, a Protein Domain with a Central β-Sheet and Two Hydrophobic Cores†. *Biochemistry* **1998**, *37*, 14508−14518.
(35) Fazelinia, H.; Xu, M.; Cheng, H.; Roder, H. Ultrafast Hydrogen Exchange Reveals Specific Structural Events during the Initial Stages of Folding of Cytochromec. *J. Am. Chem. Soc.* **2014**, *136*, 733−740.
(36) Englander, S. W.; Calhoun, D. B.; Englander, J. J.; Kallenbach, N. R.; Liem, R. K.; Malin, E. L.; Mandal, C.; Rogero, J. R. Individual breathing reactions measured in hemoglobin by hydrogen exchange methods. *Biophys. J.* **1980**, *32*, 577−589.
(37) Huang, R. Y.-C.; Chen, G. Higher order structure characterization of protein therapeutics by hydrogen/deuterium exchange mass spectrometry. *Anal. Bioanal. Chem.* **2014**, *406*, 6541−6558.
(38) Huang, L.; So, P. K.; Yao, Z. P. Protein dynamics revealed by hydrogen/deuterium exchange mass spectrometry: Correlation between experiments and simulation. *Rapid Commun. Mass Spectrom.* **2019**, *33*, 83−89.
(39) Devaurs, D.; Antunes, D. A.; Papanastasiou, M.; Moll, M.; Ricklin, D.; Lambris, J. D.; Kavraki, L. E. Coarse-Grained Conformational Sampling of Protein Structure Improves the Fit to Experimental Hydrogen-Exchange Data. *Front. Mol. Biosci.* **2017**, *4*, 13.
(40) Brodie, N. I.; Popov, K. I.; Petrotchenko, E. V.; Dokholyan, N. V.; Borchers, C. H. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Sci. Adv.* **2017**, *3*, No. e1700479.
(41) McAllister, R. G.; Konermann, L. Challenges in the Interpretation of Protein H/D Exchange Data: A Molecular Dynamics Simulation Perspective. *Biochemistry* **2015**, *54*, 2683−2692.
(42) Schenk, E. R.; Almeida, R.; Miksovska, J.; Ridgeway, M. E.; Park, M. A.; Fernandez-Lima, F. Kinetic Intermediates of Holo- and Apo-Myoglobin Studied Using HDX-TIMS-MS and Molecular Dynamic Simulations. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 555−563.
(43) Roberts, V. A.; Pique, M. E.; Hsu, S.; Li, S. Combining H/D Exchange Mass Spectrometry and Computational Docking To Derive the Structure of Protein-Protein Complexes. *Biochemistry* **2017**, *56*, 6329−6342.
(44) Zhang, M. M.; Beno, B. R.; Huang, R. Y.-C.; Adhikari, J.; Deyanova, E. G.; Li, J.; Chen, G.; Gross, M. L. An Integrated Approach for Determining a Protein-Protein Binding Interface in Solution and an Evaluation of Hydrogen-Deuterium Exchange Kinetics for Adjudicating Candidate Docking Models. *Anal. Chem.* **2019**, *91*, 15709−15717.
(45) Borysik, A. J. Simulated Isotope Exchange Patterns Enable Protein Structure Determination. *Angew. Chem., Int. Ed.* **2017**, *56*, 9396−9399.
(46) Mohammadiarani, H.; Shaw, V. S.; Neubig, R. R.; Vashisth, H. Interpreting Hydrogen-Deuterium Exchange Events in Proteins Using Atomistic Simulations: Case Studies on Regulators of G-Protein Signaling Proteins. *J. Phys. Chem. B* **2018**, *122*, 9314−9323.
(47) Zhang, Y.; Majumder, E. L.-W.; Yue, H.; Blankenship, R. E.; Gross, M. L. Structural Analysis of Diheme Cytochrome c by Hydrogen-Deuterium Exchange Mass Spectrometry and Homology Modeling. *Biochemistry* **2014**, *53*, 5619−5630.
(48) DeGrado, W. F.; Summa, C. M.; Pavone, V.; Nastri, F.; Lombardi, A. De Novo Design and Structural Characterization of Proteins and Metalloproteins. *Annu. Rev. Biochem.* **1999**, *68*, 779−819.
(49) Polizzi, N. F.; Wu, Y.; Lemmin, T.; Maxwell, A. M.; Zhang, S.-Q.; Rawson, J.; Beratan, D. N.; Therien, M. J.; DeGrado, W. F. De novo design of a hyperstable non-natural protein-ligand complex with sub-Å accuracy. *Nat. Chem.* **2017**, *9*, 1157−1164.
(50) Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203−204.
(51) Karakaş, M.; Woetzel, N.; Staritzbichler, R.; Alexander, N.; Weiner, BE; Meiler, J. BCL::Fold–de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PloS One* **2012**, *7*, No. e49240.
(52) Yang, J.; Zhang, Y. Protein Structure and Function Prediction Using I-TASSER. *Curr. Protoc. Bioinf.* **2015**, *52*, 5.8.1−5.8.15.
(53) Sripakdeevong, P.; Kladwang, W.; Das, R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 20573.
(54) Sengupta, A.; Wu, J.; Seffernick, J. T.; Sabag-Daigle, A.; Thomsen, N.; Chen, T.-H.; Capua, A. D.; Bell, C. E.; Ahmer, B. M. M.; Lindert, S.; Wysocki, V. H.; Gopalan, V. Integrated Use of Biochemical, Native Mass Spectrometry, Computational, and Genome-Editing Methods to Elucidate the Mechanism of a deglycase. *J. Mol. Biol.* **2019**, *431*, 4497−4513.
(55) Hirst, S. J.; Alexander, N.; McHaourab, H. S.; Meiler, J. RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *J. Struct. Biol.* **2011**, *173*, 506−514.
(56) Kuenze, G.; Meiler, J. Protein structure prediction using sparse NOE and RDC restraints with Rosetta in CASP13. *Proteins* **2019**, *87*, 1341−1350.
(57) Das, R.; Baker, D. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363−382.
(58) Kaufmann, K. W.; Lemmon, G. H.; Deluca, S. L.; Sheehan, J. H.; Meiler, J. Practically Useful: What the RosettaProtein Modeling Suite Can Do for You. *Biochemistry* **2010**, *49*, 2987−2998.
(59) Bender, B. J.; Cisneros, A., 3rd; Duran, A. M.; Finn, J. A.; Fu, D.; Lokits, A. D.; Mueller, B. K.; Sangha, A. K.; Sauer, M. F.; Sevy, A. M.; Sliwoski, G.; Sheehan, J. H.; DiMaio, F.; Meiler, J.; Moretti, R. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* **2016**, *55*, 4748−4763.
(60) Lindert, S.; Meiler, J.; McCammon, J. A. Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol to Improve Model Quality. *J. Chem. Theory Comput.* **2013**, *9*, 3843−3847.
(61) Frenz, B.; Walls, A. C.; Egelman, E. H.; Veesler, D.; DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* **2017**, *14*, 797−800.
(62) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P. S.; Huber, T.; Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.;

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629

Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D. A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, R. Y.; Watkins, A.; Zimmerman, L.; Bonneau, R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **2020**, *17*, 665.

(63) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031−3048.

(64) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898−7936.

(65) Pancsa, R.; Varadi, M.; Tompa, P.; Vranken, W. F. Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res.* **2016**, *44*, D429−D434.

(66) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **1999**, *34*, 82−95.

(67) Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E. M.; Baker, D. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins* **2001**, *45*, 119−126.

(68) Bonneau, R.; Strauss, C. E. M.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmström, L.; Robertson, T.; Baker, D. De Novo Prediction of Three-dimensional Structures for Major Protein Families. *J. Mol. Biol.* **2002**, *322*, 65−78.

(69) Bradley, P.; Misura, K. M. S.; Baker, D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science* **2005**, *309*, 1868.

(70) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V.; Baker, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **2009**, *77*, 89−99.

(71) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209−225.

(72) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32*, W526−W531.

(73) *The PyMOL Molecular Graphics System*, version 2.0; Schrödinger, LLC.

(74) Kim, S. S.; Seffernick, J. T.; Lindert, S. Accurately Predicting Disordered Regions of Proteins Using Rosetta ResidueDisorder Application. *J. Phys. Chem. B* **2018**, *122*, 3920−3930.

(75) Seffernick, J. T.; Ren, H.; Kim, S. S.; Lindert, S. Measuring Intrinsic Disorder and Tracking Conformational Transitions Using Rosetta ResidueDisorder. *J. Phys. Chem. B* **2019**, *123*, 7103−7112.

(76) Nguyen, D.; Mayne, L.; Phillips, M. C.; Walter Englander, S. Reference Parameters for Protein Hydrogen Exchange Rates. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 1936−1939.

(77) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V. S. R. K.; Kaas, Q.; Eletsky, A.; Huang, P.-S.; Johnsen, W. A.; Greisen, P. J., Jr.; Rocklin, G. J.; Song, Y.; Linsky, T. W.; Watkins, A.; Rettie, S. A.; Xu, X.; Carter, L. P.; Bonneau, R.; Olson, J. M.; Coutsias, E.; Correnti, C. E.; Szyperski, T.; Craik, D. J.; Baker, D. Accurate de novo design of hyperstable constrained peptides. *Nature* **2016**, *538*, 329−335.

(78) Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. Primary structure effects on peptide group hydrogen exchange. *Proteins* **1993**, *17*, 75−86.

(79) Pancsa, R.; Fuxreiter, M. Interactions via intrinsically disordered regions: What kind of motifs? *IUBMB Life* **2012**, *64*, 513−520.

(80) Uversky, V. N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114*, 6557−6560.

(81) Franzosa, E. A.; Xia, Y. Independent effects of protein core size and expression on residue-level structure-evolution relationships. *PloS One* **2012**, *7*, No. e46602.

(82) Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.* **2006**, *61*, 966−988.

(83) Farrell, D.; Anishchanka, I.; Shakeel, S.; Lauko, A.; Passmore, L. A.; Baker, D.; DiMaio, F., Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. **2020**, bioRxiv:2020.05.01.072751.

(84) Delhommel, F.; Gabel, F.; Sattler, M. Current approaches for integrating solution NMR spectroscopy and small-angle scattering to study the structure and dynamics of biomolecular complexes. *J. Mol. Biol.* **2020**, *432*, 2890−2912.

(85) Center, Ohio Supercomputer. *Ohio Supercomputer Center*, 1987.

https://doi.org/10.1021/acs.jctc.1c00077
J. Chem. Theory Comput. 2021, 17, 2619−2629