Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data

Melanie L. Aprahamian[1], Emily E. Chea[2], Lisa M. Jones[2], Steffen Lindert[1,]*

[1] Department of Chemistry and Biochemistry, Ohio State University, Columbus, OH, 43210

[2] Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD, 21201

Key words: Rosetta, computational protein structure prediction, ab initio protein structure prediction, mass spectrometry covalent labeling, hydroxyl radical footprinting, FPOP

Abstract

In recent years mass spectrometry-based covalent labeling techniques such as hydroxyl radical footprinting (HRF) have emerged as valuable structural biology techniques yielding information on protein tertiary structure. This data, however, is not sufficient to predict protein structure unambiguously, as it only provides information on the relative solvent exposure of certain residues. Despite some recent advances, no software currently exists that can utilize covalent labeling mass spectrometry data to predict protein tertiary structure. We have developed the first such tool, which incorporates mass spectrometry derived protection factors from HRF labeling as a new centroid score term for the Rosetta scoring function to improve the prediction of protein tertiary structure. We tested our method on a set of four soluble benchmark proteins with known crystal structures and either published HRF experimental results or internally acquired data. Using the HRF labeling data, we rescored large decoy sets of structures predicted with Rosetta for each of the four benchmark proteins. As a result, the model quality improved for all benchmark proteins, as compared to when scored with Rosetta alone. For two of the four proteins, we were even able to identify atomic resolution models with the addition of HRF data.

Introduction

Historically, mass spectrometry has been used as a tool to quantify the mass and oligomeric distribution of proteins and protein assemblies.[1, 2] More recently, advances have been made that allow mass spectrometry experiments to yield three-dimensional structural information on proteins and their complexes. By itself, there is no one mass spectrometry technique that can unambiguously elucidate atomic-resolution tertiary structure of a protein or protein complex. Hence, a combination of multiple different techniques is generally required.[3-5] Several techniques

have been particularly successful in probing the tertiary structure of proteins and their complexes. Hydrogen-deuterium exchange (HD/X) is based upon measuring the extent of isotopic exchange of backbone amide hydrogens.[6, 7] Chemical cross-linking involves studying the structurally defined distances by covalently pairing functional groups within a protein.[8, 9] Non-covalent interactions between lysine residues and 18-crown-6 ether (a cyclic organic compound) can provide lysine solvent accessibility within proteins.[10] Finally, covalent labeling (sometimes referred to as "protein footprinting") involves exposing a protein in solution to a small labeling reagent that will covalently bond to select amino acid sidechains that are exposed to solvent, whereas sidechains buried within the core of the protein or occluded by interacting protein subunits will not get labeled.[11-13] This provides information about the relative location of certain amino acids with respect to the solvent (either on the surface and solvent exposed or buried within the protein or protein complex structure). A variety of different labeling reagents exist and some are highly specific as to which amino acid(s) can react with the reagent and others have a much broader range of potential target residues. These techniques have been successfully employed with mass spectrometry to analyze protein structures.[14-22]

One covalent labeling method which recently has been increasingly widely used is hydroxyl radical footprinting (HRF).[23, 24] This method involves exposing a solvated protein of interest to hydroxyl radicals generated from one of a variety of sources. Initially, oxidative labeling was performed using a synchrotron that ionized water to form the hydroxyl radicals.[25] With recent advancements, a new method of hydroxyl radical labeling, fast photochemical oxidation of proteins (FPOP), has been developed.[26, 27] With FPOP, a pulsed laser is used to photolyze hydrogen peroxide on a microsecond timescale, which is faster than the unfolding of a protein. This ensures that the labeling process does not disrupt the native state of the protein. In conjunction with mass

spectrometry, FPOP provides important insight into the structure of proteins. This labeling method is quite broad in that it can label 19 of the 20 different amino acids, yielding extensive structural information. Despite the wealth of information provided by FPOP, the data itself is sparse, meaning that the solvent exposure information of a set of protein residues cannot provide unambiguous determination of protein structure. There remains a critical need for computational methods that can facilitate and compliment the structural interpretation of mass spectrometry FPOP labeling data.

Over the years, numerous experimental techniques have been successfully combined with computational methods to predict protein structures. Some examples of this are sparse experimental data from site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) in conjunction with Rosetta to improve protein structure predictions,[28, 29] nuclear magnetic resonance spectroscopy (NMR),[30, 31] small-angle X-ray scattering (SAXS),[32-35] and cyro-electron microscopy (cryo-EM).[36-43] Mass spectrometry techniques have also been utilized in conjunction with computational methods. Malmström and coworkers have made significant contributions by incorporating data from MS chemical cross-linking experiments as inputs into computational methods for protein structure prediction.[15, 44-47] The work of Sali and coworkers has contributed greatly to the field with the development of the Integrative Modeling Platform (IMP), an open source platform that integrates experimental data into computational methods.[19, 35, 48-52] IMP is designed as a set of self-contained modules that can be mixed and matched based upon a user's preference. Models are generated and scored based upon spatial restraints that are derived from multiple sources of experimental data. Currently IMP supports the use of experimental data gathered from sources such as SAXS profiles, EM images and density maps, NMR, chemical cross linking, HD/X, and chromosome conformation capture. With IMP, both monomeric and multi-

unit protein structures can be studied. Finally, Yang and coworkers have developed an integrative method, iSPOT, to determine protein-protein complexes that combines SAXS, hydroxyl radical footprinting, and computational docking of either rigid-body or molecular dynamics models.[32]

Computational modeling using FPOP data is still in its early stages. Recently, an integrated workflow was developed by Xie and coworkers that successfully demonstrated correlation between experimental high-resolution hydroxyl radical footprinting data and residue solvent exposure (as measured by absolute average solvent accessible surface area) as well as differentiated between low and high RMSD models for the soluble proteins myoglobin and lysozyme.[53] This elegant work demonstrated that there is strong potential for successfully incorporating HRF or FPOP experimental data into computational methods in order to improve protein structure prediction. Despite the many advances and successes with using sparse data from various experimental methods for structure prediction, the use of covalent labeling mass spectrometry as the data source had yet to be accomplished.

In this work, we incorporated mass spectrometry derived protection factors from FPOP and synchrotron-based HRF labeling as a new score term for the Rosetta scoring function to improve the prediction of protein tertiary structure. Rosetta is one of the primary computational tools used for protein structure prediction.[54] To accomplish our goal, we compiled a set of four soluble benchmark proteins with known crystal structures and either published HRF/FPOP experimental results or internally acquired data. We developed an efficient metric to quantify residue-specific burial that correlated linearly to the natural logarithm of experimental protection factors derived from the labeling rates. A new Rosetta centroid score term, that utilizes residue-resolved protection factors as inputs, was developed. This score term was used in conjunction with the standard Rosetta scoring function to rescore large decoy sets of predicted structures for each of the four benchmark

proteins. In this process of rescoring, the quality of all models improved such that after rescoring the structures with the best score correlated more closely to the native structures. For two of the four proteins, we were even able to identify atomic resolution models using the HRF/FPOP data.

Materials and Methods

Benchmark Dataset and Experimental Protection Factors

For this work, we used protection factor (PF) which was first described by Chance and coworkers and is derived from a labeling rate constant as a metric for residue labeling.[55] PF is defined as the relative intrinsic reactivity of a given residue to hydroxyl radicals divided by the rate constant. The intrinsic reactivities of each amino acid type are well defined in the literature.[24] The PF, as expressed on a natural logarithmic scale, has been shown to correlate with the solvent exposure of a given residue.[16, 55, 56] Within the literature, the PF has been defined multiple ways, but for our purposes we have defined the protection factor for residue i, where R_i is the intrinsic reactivity for residue i and k_i is the experimentally determined labeling rate constant, as defined by eq 1:

$$PF_i = \frac{R_i}{k_i} \tag{1}$$

As a benchmark set, four different proteins with available FPOP or HRF labeling data were utilized. These proteins were calmodulin (PDB ID: 1PWR), myoglobin (PDB ID: 1DWR), lysozyme (PDB ID: 1DPX), and cytochrome c (PDB ID: 2B4Z). The experimentally determined PFs for calmodulin were extracted from the published work of Kaur et al who generated radicals via a millisecond timescale synchrotron radiation method.[16] For myoglobin, the PFs were calculated from the reported labeling rate constants by Xie et al.[53] using the reactivities reported in the literature.[24] For this study, radicals were generated using sub-microsecond FPOP with a

dosimeter to provide varying doses of radicals. Finally, the experimental PFs for both lysozyme and cytochrome c were oxidatively modified by FPOP at a single radical dose as described in the Supporting Information.

For incorporation of the data into the newly developed score term, input files were created for each protein consisting of a heading line followed by two columns comprising the residue number and the natural logarithm of the protection factor, with each labeled residue on a new line. FPOP/HRF can label 19 of the 20 amino acids, however data from the following residue types were omitted due to either too low/high reactivity or unclear products: M, C, D, N, Q, T, S, A, G, R, K, and V. Of this list of omitted residues, it has been previously suggested by Xie et al. that the sequence context plays a role in whether or not these amino acid types are labeled. This is a complex issue and has not been examined in this current work. As a result, only eight of the twenty amino acids were considered in the analysis: I, L, P, F, W, Y, E, and H. These residues have intermediate reactivities and correspond with the residue types utilized in similar studies.[16, 53]

Rosetta ab initio Folding

In the absence of any experimental labeling data, decoy sets of 20,000 structures were generated for each of the four benchmark proteins using the *AbinitioRelax* application within Rosetta.[57-59] The *AbinitioRelax* protocol consists of two main steps: 1) a coarse-grained fragment-based search of conformational space that uses a low-resolution "centroid"-based (treating each residue with backbone atoms defined explicitly and the side-chain represented as a single sphere) scoring function and 2) a high-resolution refinement using the full-atom Rosetta score function.

The generated decoy sets act as benchmarks to compare the structure prediction capabilities of Rosetta in the absence of FPOP/HRF labeling data. Specifics of the protocol have been detailed extensively in the literature.[60] The fragment libraries for this work were generated using the

Robetta online server.[61] The required FASTA formatted sequences and native protein structures were extracted from each protein's respective PDB file. The fragment libraries, FASTA sequences, and native PDB structures (used solely for determining the deviation of the generated models from the native) were used as inputs for Rosetta's *AbinitioRelax* application. For lysozyme, disulfide bonds were present between the following residues: 6 and 127, 30 and 115, 64 and 80, and 76 and 94. An additional input file was provided to specify the residues that are a part of the disulfide bonds. The generated structures were scored using the Rosetta energy function (Ref15), where the score is an approximation of the energy of the protein or complex.[62] The scores and respective root mean square deviation (RMSD) to the native crystal structure were extracted from the output score file. Structures were ranked based upon their scores with lower scores anticipated to correspond to models closer in structure to the native. Rosetta score versus RMSD to the native protein were generated to demonstrate this correlation.

For each of the benchmark proteins, two small sets of representative structures were generated. The first set represented ten native-like conformations of each protein which were obtained by relaxing each crystal PDB in the Rosetta force field using the *relax* application.[63, 64] We will refer to these structures as the ten native-like models or the native-like model set. The second set contained models that scored well with the Rosetta energy function, but had high RMSDs compared to the crystal native structures. These were obtained by extracting the top ten scoring models with RMSD > 10 Å for each protein from the initial *ab initio* calculations. We will refer to these structures as the good scoring/high RMSD model set. Together, these sets represented the two extremes of potential models that we desired to efficiently differentiate between using our new score term.

Residue Exposure Metric

To compare the protection factors extracted from the FPOP/HRF labeling data to residue exposure in protein models, a corresponding residue exposure measure was developed which enabled calculation of the level of exposure of every labeled residue in a protein model. The PF has been shown to correlate to a residue-level solvent accessible surface area (SASA).[16, 53, 56] Because residue-level SASAs are expensive to calculate,[65, 66] we explored other metrics, aside from SASA, that were less computationally expensive and provided even stronger correlation to the natural logarithm of the experimental FPOP/HRF PFs. Assuming solvent exposed residues are preferentially labeled, we sought to find a residue burial/exposure metric that showed correlation to the natural logarithm of the PFs. Several methods, such as weighted neighbor count and SASA,[65, 67] were investigated. For reference, the correlation between SASA and the natural logarithm of the PFs can be found in Supplemental Figure S-1. However the burial measure found to give the strongest correlation to the experimental data was a neighbor count determined for each labeled residue. A residue with a high neighbor count can be thought of as buried whereas a residue with a low neighbor count can be considered solvent exposed. For this analysis, a low-resolution model of the protein was used where all of the backbone atoms were represented explicitly and the side-chain was represented as a single sphere called a centroid. To calculate a residue's neighbor count, the distances between the labeled residue's centroid (residue i) and all other residues' centroids (residues $j \neq i$) were measured. The distance, r_{ij}, was then used in a sigmoid function that defined a value between 0 and 0.7, as shown in Supplemental Figure S-2, representing the amount of contribution of a neighboring residue *j* to the total neighbor count of the target residue i. The closer a residue j's centroid is to labeled residue i's centroid, the more it contributed to the overall neighbor count; conversely, the further away it is, the less it contributed. The total neighbor

count for each labeled residue i was then defined as the sum of every residue's contribution to the neighbor count:

$$neighbor\ count_i = \sum_{j \neq i}^{total\ \#\ residues} \frac{1.0}{1.0 + e^{0.1(r_j - 9.0)}} \quad (2)$$

We developed a new Rosetta application, *burial_measure_centroid*, which calculated the neighbor counts (as defined in eq 2) for arbitrary protein structures. For each of the eighty models comprising the native-like and good score/high RMSD model sets, the neighbor counts were calculated using the *burial_measure_centroid* Rosetta application. The neighbor counts for the ten native-like structures of calmodulin (1PRW) were used to perform a linear regression with the corresponding experimental lnPF values. The linear fit obtained was then used as a prediction function to predict the neighbor count for all 80 representative models with their respective experimental lnPF values as inputs.

hrf_ms_labeling Score Term

A new score term, *hrf_ms_labeling*, was developed to be incorporated into Rosetta to assess the agreement of Rosetta models with experimental FPOP/HRF labeling data. This score term was defined as a centroid score term that rewards protein conformations that show agreement with the experimental labeling data. By treating the score term in a Bayesian fashion, the total Rosetta score was derived (as shown explicitly in the Supporting Information) to be the sum of the weighed score term and the current Rosetta score:

$$Total\ Score = w_{\mathrm{hrf}} * hrf\_ms\_labeling + RosettaScore \quad (3)$$

The score term, *hrf_ms_labeling*, was implemented using the linear prediction function obtained by correlating the observed neighbor counts and experimental lnPF for the benchmark protein calmodulin (see the previous section, Residue Exposure Metric). A value for *hrf_ms_labeling* was calculated by summing the per-residue neighbor scores over the set of labeled residues and was defined as:

$$hrf\_ms\_labeling = \sum_{i}^{\#\text{ labeled residues}} \frac{-1.0}{1.0 + e^{2.0(|diff|_i - 7.5)}} \tag{4}$$

where $|diff|_i$ is the absolute value of the difference between the observed neighbor count (calculated using eq 2 for the modeled protein) and the predicted neighbor count (calculated using the linear prediction function) for labeled residue i. Using the definition in eq 4, each labeled residue contributed a per-residue score ranging from -1 to 0 with a value of -1 in case of strong agreement with the experiment and a value of 0 in case of complete disagreement. If the value of $|diff|_i$ fell between 5 and 10 (which corresponded to the same cutoffs as the delta lines used in analyzing the prediction function), the residue received a logistically increasing value ranging from -1 to 0. The per-residue score (function found within the summation in eq 4) is depicted in Figure 1 with all relevant points highlighted.

Rescoring of Rosetta Structures

To test the capability of our new score term to improve Rosetta model quality, the 20,000 Rosetta models initially generated as part of the *ab initio* folding for each benchmark protein were rescored with the *hrf_ms_labeling* score term. The calculated *hrf_ms_labeling* score was weighted by a value of 6.0 and added to the Rosetta score calculated using Rosetta's Ref15 energy function:

$$Total\ Rosetta\ Score = Ref15\ Rosetta\ Score + 6.0 * hrf\_ms\_labeling$$

(5)

A weight of 6.0 was the lowest possible value that showed the greatest improvement. We iterated through all integer values from 1-36 and determined the top scoring models' RMSDs at each weight. The results of this analysis are shown in Supplemental Figure S-3. To calculate the *hrf_ms_labeling* contribution for each model, the *score* Rosetta application was run on each of the 80,000 models using the output structures from the initial *ab initio* model generation as input. For each of the 80,000 rescored models, the total Rosetta scores, the RMSD to the native structure, and the *hrf_ms_labeling* scores were extracted.

Model Evaluation

Several different metrics were used to evaluate the performance of both Rosetta and the score term. Those metrics were based upon the concept of an energy funnel, i.e. that within the overall energy landscape, low RMSD models can be distinguished from high RMSD models due to their lower energy (Rosetta score).[68] The first metric used was a simple determination of the top scoring model's RMSD to the native structure. In practice, the Rosetta model with the lowest (most favorable) Rosetta score is assumed to be closest in structure to the native. Because all the benchmark proteins chosen for this study had crystal PDB structures available, an RMSD for that model can be calculated.

The second metric used was the goodness-of-energy-funnel metric P_{near}, as defined by Bhardwaj et al.[69] A value of P_{near} was calculated for each Rosetta score versus RMSD distribution using the following equation:

$$P_{\text{near}} = \frac{\sum_{m=1}^{N} exp\left(-\frac{rmsd_m^2}{\lambda^2}\right) exp\left(-\frac{E_m}{k_B T}\right)}{\sum_{m=1}^{N} exp\left(-\frac{E_m}{k_B T}\right)}$$

(6)

where N is the total number of models and E_m and $rmsd_m$ are the Rosetta score and RMSD of model m. The parameter λ was given a value of 2.0 and controlled how similar a model needed to be to the native to be considered native-like. The final parameter, k_BT, was set to 1.0 and governed the shallowness or depth of the funnel affects P_{near}. Values of P_{near} can range from 0 (very non-funnel like) to 1 (funnel-like).

The final metric used was a comparison of the number of top 100 scoring models with RMSD's below a 10.0 Å. By comparing this metric between different Rosetta score versus RMSD distributions we were able to investigate how well (or poorly) the addition of *hrf_ms_labeling* was at improving model quality.

Results & Discussion

Generation of Control ab initio Model Set for Benchmark Proteins using Rosetta

To establish the baseline performance of Rosetta's Ref15 scoring function at predicting protein structures without any additional experimental knowledge, decoy sets consisting of 20,000 models were generated for each of four benchmark proteins. The four proteins selected for the benchmark were calmodulin (PDB ID: 1PWR), myoglobin (PDB ID: 1DWR), lysozyme (PDB ID: 1DPX), and cytochrome c (PDB ID: 2B4Z). Table 1 summarizes the benchmark proteins. These proteins ranged in size from 104 to 153 amino acids in length. Contact orders (CO) were calculated for each of the proteins.[70] The contact orders for all four proteins were low, ranging from 10.7 to 13.7. The secondary structure content for the four proteins were relatively high, ranging from 41% to 74%. Because these proteins were all relatively small (approx. fewer than 150 amino acids), had high secondary structure content and low contact orders, we concluded that they were amendable to Rosetta *ab initio* protein structure prediction.

Using Rosetta to generate 20,000 models for each of the four proteins resulted in the selection of best-scoring structures with RMSDs ranging from 5.0 Å to 15.2 Å, as summarized in Table 2 and indicated on the Rosetta score versus RMSD to native structure plots in panel A of Figure 2 by stars. The two proteins with top scoring structures that were closest to their respective native structures were myoglobin (RMSD = 5.0 Å) and cytochrome c (RMSD = 5.5 Å). The predictions for the remaining two proteins, calmodulin and lysozyme, were poor, yielding top scoring models with RMSD's of 11.8 and 15.2 Å, respectively. Considering the size of the benchmark proteins, none of these best-scoring models were high-quality, near-atomic resolution models. For two of the proteins, even an incorrect topology was identified. However, as can be seen in Figure 2A, models with significantly lower RMSDs to the native structure were built for all four proteins. For calmodulin, the RMSDs for the generated models ranged from 2.9 Å to 21.5 Å. Similar ranges were sampled for cytochrome c and myoglobin, with RMSDs ranging from 1.4 Å to 21.3 Å and 1.5 Å to 27.3 Å, respectively. Lysozyme had the poorest sampling, where model RMSDs ranged from 6.0 Å to 18.7 Å. This indicated that better, and in some cases even near-atomic resolution models, were in fact generated for all proteins, but they were generally not identified by the lowest score.

The goodness-of-energy-funnel metric, P_{near}, was used to evaluate the funnel quality of each of the distributions. As can be seen in Table 2, none of the distributions had P_{near} values greater than 0.1, strongly suggesting that none of the ensembles of models exhibited funnel-like score distributions. This lack of a funnel in the Rosetta score versus RMSD to native structure plots made structure prediction and particularly native structure identification challenging. Based upon these *ab initio* structure prediction results, we concluded that incorporation of experimental data,

such as HRF/FPOP labeling data, had the potential to improve identification of low RMSD models by score.

Rescoring Model Sets using hrf_ms_labeling

The overall goal of this work was to utilize experimental HRF/FPOP labeling data in order to improve models predicted by Rosetta. To accomplish this, a new Rosetta score term, *hrf_ms_labeling,* was developed that incorporated experimental HRF/FPOP protection factors (PFs). After developing *hrf_ms_labeling,* we confirmed that incorporation of HRF/FPOP labeling data did enable discrimination of near-native and high RMSD models and that combination of this score with the total Rosetta Ref15 score did improve the quality of the models selected from the structure ensembles.

The first step in this process was to demonstrate that a correlation existed between the experimental labeling data (the PFs) and a residue solvent exposure metric derived within Rosetta. The metric that demonstrated the best correlation was the per-residue neighbor count, as defined in the Methods section. The calculated neighbor count for every labeled residue within calmodulin (1PRW), one of our benchmark proteins, was plotted against the natural logarithm of the respective PF values. The positive correlation, as seen in Figure 3, had an R^2 of 0.48 and p-value of 1.36E-36. The observed trend matched our expectation where residues with a low lnPF also showed a low neighbor count (suggesting a higher solvent exposure) and residues with a high lnPF showed a high neighbor count (suggesting a lower solvent exposure). The derived relationship between PFs and neighbor count was used to predict neighbor counts for all four benchmark proteins based on the experimental HRF/FPOP protection factors. For comparison, observed neighbor counts for two small sets of representative structures (the native-like model sets and the good scoring/high RMSD model sets) were calculated from each pdb structure using *burial_measure_centroid*. The

predicted neighbor counts have been plotted against the observed neighbor counts (calculated directly from representative structures of the four benchmark proteins) in Figure 4. In order to quantify the accuracy of the prediction, two delta lines were defined ($d_1 = 5.0$ and $d_2 = 10.0$). These delta lines represent how close the predicted neighbor counts were to the actual observed values. Using the native-like model sets for all four proteins, an average of 81% and 59% of the labeled residues fell within d_2 and d_1, respectively, whereas only 67% and 38% of those belonging to the good scoring/high RMSD model sets did. This demonstrated that we predicted the majority of the labeled residues in native-like models within the delta lines and simultaneously excluded the majority of residues in the high RMSD models from within the delta lines. This suggested that agreement between a model's residue exposure and the neighbor count metric derived from experimental FPOP/HRF mass spectrometry data can indeed distinguish between low and high RMSD models and can thus be used to rescore protein models built in the absence of experimental FPOP/HRF labeling data. To be able to rescore protein models, a *hrf_ms_labeling* score term was developed for incorporation into Rosetta.

We next demonstrated that the new score term was effective in improving model prediction. The 20,000 model decoy sets generated for each of the four benchmark proteins were rescored with the *hrf_ms_labeling* term added to the Ref15 Rosetta score. For each set of models, Rosetta score + *hrf_ms_labeling* versus RMSD plots were generated. Based upon the rescored structures, new top scoring models were selected. As shown in Table 3, the RMSDs of the top scoring models improved for all four proteins, while for two of the proteins near-atomic resolution models were identified. The biggest increases in top scoring model quality were observed for lysozyme. Addition of HRF/FPOP labeling data improved the RMSD of the top scoring lysozyme model from 15.2 Å to 7.2 Å, a significant improvement in the model's quality. Although a model with

an RMSD of 7.2 Å is not usually considered high quality, considering that the best lysozyme *ab initio* model had an RMSD of 6.0 Å, one of the best existing models was identified. Both myoglobin and cytochrome c showed decreases in their RMSDs to near-atomic resolution models (2.2 and 1.8 Å respectively), also identifying models with RMSDs close to the best existing models within the 20,000 structures. Calmodulin had the least improvement with a change in RMSD from only 11.8 to 10.2 Å. When superimposing the top scoring models onto their respective native structures, as depicted in panels B and D of Figure 2, a significant increase in model quality could be observed as a result of the addition of *hrf_ms_labeling*. All top scoring models now identify the correct protein topology.

In addition to analyzing the RMSD of the top scoring models, the overall energy landscape of the structures was analyzed. Values of P_{near} were calculated for each score versus RMSD distribution, identical to what was done without the addition of *hrf_ms_labeling* (see Table 2). With the addition of the *hrf_ms_labeling* term to the scoring function, there was an increase in P_{near}, i.e. an increase in funnel quality of the score vs RMSD plots, for all four proteins. As can be seen in panel C of Figure 2, the distributions appear more funnel like with lower RMSD models receiving lower scores. Interestingly, the P_{near} values of the two proteins for which near-atomic resolution models were identified (myoglobin and cytochrome c) were several orders of magnitude higher than those of the other proteins. We thus speculated that P_{near} might be used as a confidence measure to identify cases for which near-atomic resolution models were identified. To explore this idea we recalculated score vs RMSD plots with respect to the lowest scoring structure (to obviate the necessity for knowledge of the native structure) and measured P_{near} values for these distributions as shown in the last column of Table 2. While the trend was not as pronounced as before, this P_{near} value still served as a confidence measure in that the P_{near} values of the two

proteins for myoglobin and cytochrome c were more than two orders of magnitude higher than those of the other proteins. Upon rescoring with *hrf_ms_labeling,* the overall distribution of structures did not shift to a lower RMSD, because *hrf_ms_labeling* was simply used to rescore previously generated models. Plots of *hrf_ms_labeling* versus RMSD are shown in Supplemental Figure S-4. For all four proteins, models with poor (i.e. high, closer to 0) *hrf_ms_labeling* scores also had a higher RMSD. Likewise, some of the models with a better *hrf_ms_labeling* score tended to have a lower RMSD. There were a fair number of models however that had good *hrf_ms_labeling* scores but a high RMSD. This trend is not concerning, because the information obtained from the HRF/FPOP labeling experiments are not all encompassing of a proteins structure. Individual score terms within Rosetta generally do not exhibit the exact trend of low score/low RMSD and high score/high RMSD. Combination of this score term with the Rosetta scoring function however generated the desired trend.

We finally investigated whether a larger set of top scoring models after the rescoring were of increased quality. Histograms were generated showing the RMSD frequency of the top 100 scoring models for the distributions pre- and post-addition of *hrf_ms_labeling*. Based upon these histograms shown in Figure 5, there was a definite shift in the model quality for calmodulin and myoglobin, with more models scoring well with low RMSDs. The percentage of the top 100 scoring models that had a RMSD < 10 Å increased from 35% to 68% for calmodulin with the addition of *hrf_ms_labeling.* This illustrates that despite not identifying a near-atomic resolution model for calmodulin, addition of the labeling information significantly improved the model quality. Myoglobin demonstrated an increase in the percentage of models in the top scoring 100 with RMSD < 5 Å from 47% to 70%. A shift in model quality of the top 100 scoring models was also seen with for lysozyme and cytochrome c, albeit much less significant.

The *hrf_ms_labeling* score term has shown great success in rescoring structures based on experimental HRF/FPOP labeling data and has been designed efficiently. A centroid form of the score term was chosen for two reasons. First, this implementation showed the highest correlation between the centroid based neighbor count and experimental lnPFs. Second, a centroid-based score function is crucial in predicting structures within Rosetta's *AbinitioRelax* protocol. Within this protocol, the main sampling of conformational space occurs during the centroid scoring phase. Thus *hrf_ms_labeling* would have maximal impact on predicting structures *ab initio* if it was utilized during the centroid scoring phase. Future work will focus on developing these *ab initio* capabilities.

Conclusion

In this work, a new Rosetta score term, *hrf_ms_labeling,* was developed. This score term utilizes residue-resolved protection factors from hydroxyl radical labeling (HRF/FPOP) mass spectrometry data and assesses agreement of protein model with the experimental data. Four proteins (calmodulin, cytochrome c, myoglobin, and lysozyme) which had both available experimental data and known crystal structures were used to benchmark the performance of the score term. Using the linear correlation between the natural logarithm of the experimental protection factors and calculated neighbor counts for one of the benchmark proteins, calmodulin, a prediction function was generated to predict the neighbor counts for the other proteins using their respective lnPFs. This prediction function was used as the basis of the new score term *hrf_ms_labeling.* The new score term was used to rescore sets of 20,000 models for each protein generated using Rosetta's *AbinitioRelax* application. As a result, the top scoring model increased in quality for all four proteins. The method used for radical generation did not adversely affect the modeling. For two of the four proteins, we were even able to identify atomic resolution models

using the HRF/FPOP data. In addition, the overall distribution of models moved more towards a funnel-like energy landscape, indicating that good scoring models were closer in structure to their respective natives. Finally, we were able to identify a confidence measure that has the potential to identify successful models without having to know the native structure.

To our knowledge, we are reporting the first method to incorporate experimental HFR/FPOP labeling data in protein structure prediction. This marks an important first step in fully utilizing mass-spectrometry-based covalent labeling techniques in quantitative structure predictions, rather than just qualitative explanations. By demonstrating the potential of covalent labeling in conjunction with the protein structure prediction capabilities of Rosetta, these techniques will be elevated to be comparable in utility to other structural biology techniques such as EPR or FRET. The scoring term and applications discussed in this paper are freely available and easily accessible through Rosetta. We have added a tutorial, including a summary of necessary files and command lines to the supporting information.

Future work will focus on extending this methodology to other labeling techniques. While this particular scoring term is specific to HRF, we plan to implement the capability to use labeling data from other mass-spectrometry-based covalent labeling experiments in the future. A second direction of our future efforts will be to develop covalent labeling-guided *ab initio* structure prediction, where the labeling data is used as part of the actual structure generation as opposed to rescoring structures generated in the absence of the experimental data.

ASSOCIATED CONTENT

Supporting Information Available: Experimental methods, Bayesian derivation of *hrf_ms_labeling*, additional figures, and a tutorial for the use of the new score term in Rosetta. This information is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

* Steffen Lindert

Department of Chemistry and Biochemistry, Ohio State University

2114 Newman & Wolfrom Laboratory, 100 W. 18th Avenue, Columbus, OH 43210

614-292-8284 (office), 614-292-1685 (fax)

lindert.1@osu.edu

Funding Sources

This work was supported by NSF (CHE 1750666) to S.L. Additionally, work in the Lindert laboratory is supported through NIH (R03 AG054904, R01 HL137015), and a Falk Medical Research Trust Catalyst Award. Work in the Jones lab is supported by NSF (MCB 1701692).

ACKNOWLEDGMENTS

The authors would like to thank the members of the Lindert lab and Rosetta Commons, in particular Andrew Leaver-Fay, for many useful discussions. We would like to thank the Ohio Supercomputer Center for valuable computational resources.[71]

Figures

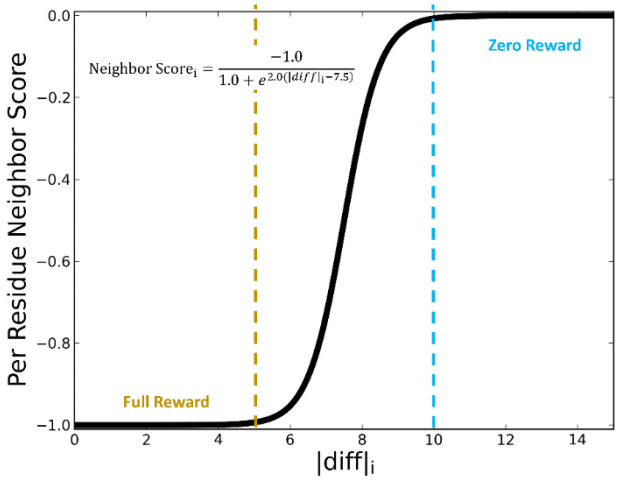

Figure 1. Plot of the per-residue neighbor score for labeled residue i as a function of the absolute difference between its observed and predicted neighbor counts ($|diff|_i$). The score function fully rewarded (with a score of -1) residues that have an $|diff|_i < 5$ and gave no reward (a score of 0) to residues that have an $|diff|_i > 10$.

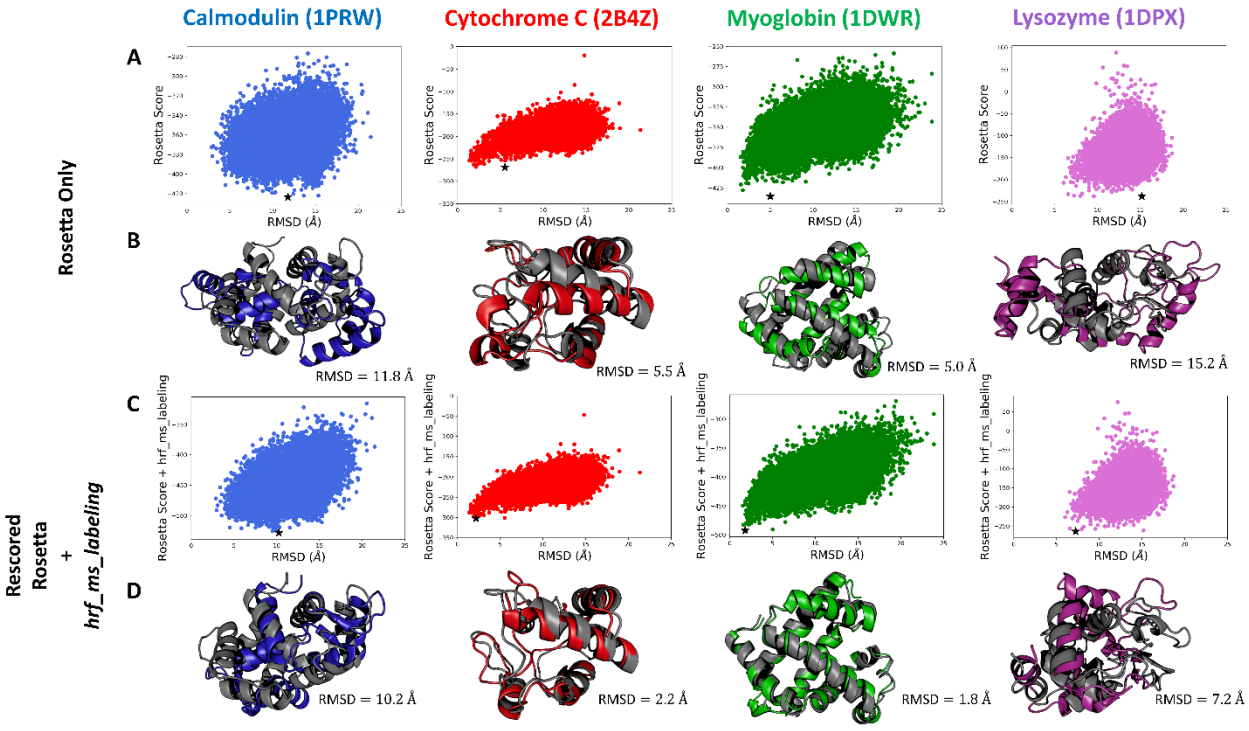

Figure 2. (A) Rosetta score versus RMSD to the native structure plots for 20,000 models generated using Rosetta *ab initio* for each of the four benchmark proteins. The top scoring model is

represented as a star on each plot. (B) The top scoring models from the Rosetta score versus RMSD distributions in A (color) superimposed upon the respective native model (grey). (C) Rosetta score + *hrf_ms_labeling* versus RMSD to the native structure plots for each of the four benchmark proteins after rescoring with the new score term. The top scoring model is represented as a star on each plot. (D) The top scoring models from the Rosetta score + *hrf_ms_labeling* rescoring distributions in C (color) superimposed upon the respective native model (grey).

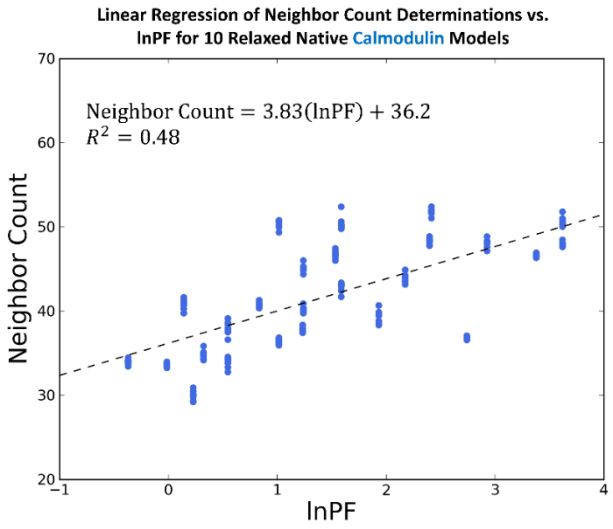

Figure 3. Linear regression between the neighbor count and the natural logarithm of the experimental protection factor (lnPF) for ten relaxed native models of calmodulin. The linear fit along with its coefficient of determination are indicated on the plot.

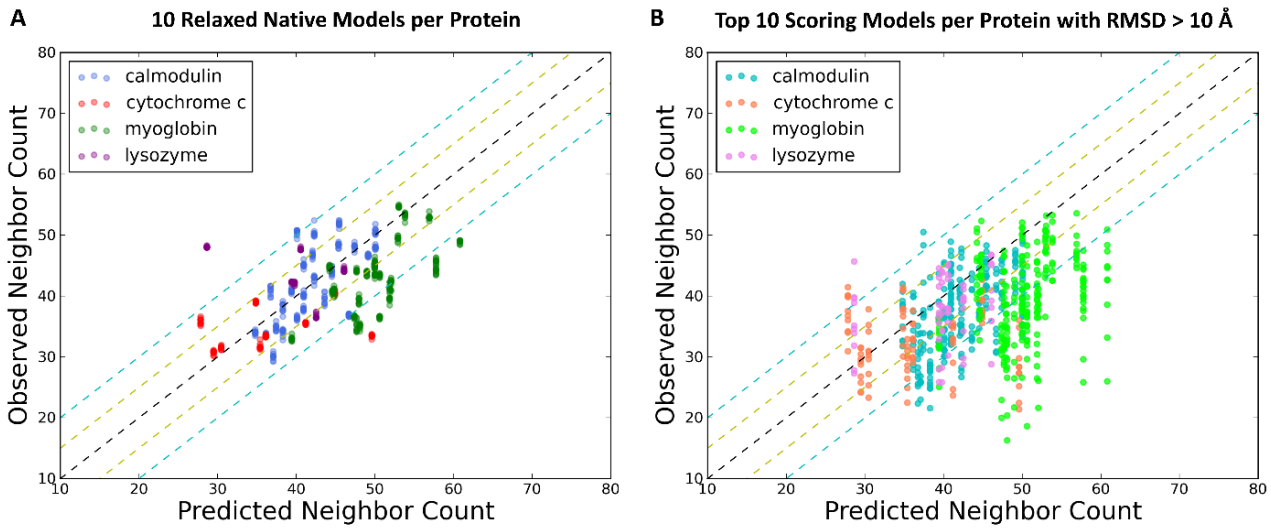

Figure 4. (A) Plot of predicted and observed neighbor counts for ten relaxed native models for each of the four benchmark proteins. (B) Plot of predicted and observed neighbor counts for ten models with good Rosetta scores and high RMSD values (> 10 Å) as compared to their respective natives for each of the four benchmark proteins. For both plots, the dashed black line represents the theoretical perfect fit (the predicted matches the observed perfectly) and the yellow and cyan lines represent the inner ($d_1 = 5$) and outer delta ($d_2 = 10$) lines respectively.

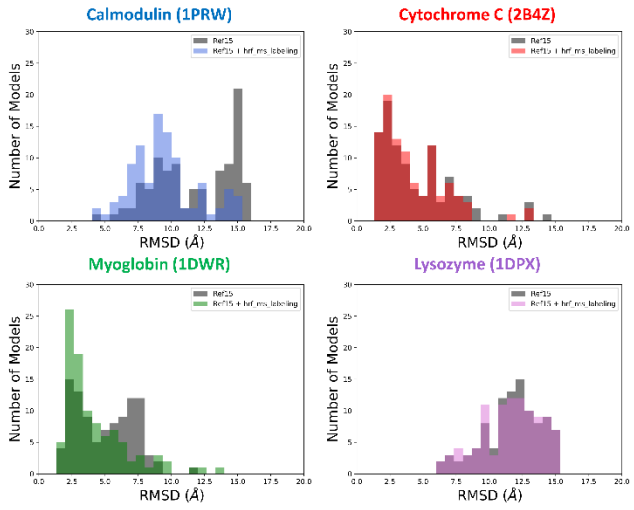

Figure 5. Histograms for each of the four benchmark proteins showing the RMSD frequency of the top 100 scoring models from both the ensembles generated using Rosetta and the ensembles

obtained after rescoring with *hrf_ms_labeling*. The histograms are plotted ranging from 0 to 20 Å with bin widths of 0.67 Å.

Tables

Table 1. Summary of the four benchmark proteins.

Protein	PDB ID	Number of Amino Acids	Number of Labeled Residues	Contact order	Secondary Structure Content (%)
Calmodulin	1PRW	148	25	10.7	61
Cytochrome C	2B4Z	104	9	11.6	41
Myoglobin	1DWR	153	25	12.4	74
Lysozyme	1DPX	129	6	13.7	51

Table 2. Rosetta ab initio prediction and rescoring results summary with and without the addition of *hrf_ms_labeling*

	Rosetta ab initio Results		Rosetta + hrf_ms_labeling Rescore Results		
Protein	Top Scoring Model RMSD to Native (Å)	P_{near}	Top Scoring Model RMSD to Native (Å)	P_{near}	Confidence Measure (P_{near} to Top Soring RMSD)
Calmodulin (1PRW)	11.8	2.10 E-8	10.2	1.17 E-6	4.18 E-5
Cytochrome C (2B4Z)	5.5	0.0805	2.2	0.238	0.038
Myoglobin (1DWR)	5.0	0.00208	1.8	0.378	0.0089
Lysozyme (1DPX)	15.2	3.04 E -7	7.2	1.89 E-6	3.079 E-9

References

1. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** *246* (4926), 64-71.

2. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R.; Iii, Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology* **1999,** *17* (7), 676.

3. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003**.

4. Küster, B.; Mann, M., Identifying proteins and post-translational modifications by mass spectrometry. *Current Opinion in Structural Biology* **1998,** *8* (3), 393-400.

5. Pi, J.; Sael, L., Mass Spectrometry Coupled Experiments and Protein Structure Modeling Methods. *International Journal of Molecular Sciences* **2013,** *14* (10), 20635-20657.

6. Zhang, Z.; Smith, D. L., Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Science : A Publication of the Protein Society* **1993,** *2* (4), 522-531.

7. Katta, V.; Chait, B. T., Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. *Rapid communications in mass spectrometry: RCM* **1991,** *5* (4), 214-217.

8. Sinz, A., Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass Spectrometry Reviews* **2006,** *25* (4), 663-682.

9. Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G., High throughput protein fold identification by using experimental

constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** *97* (11), 5802-5806.

10. Ly, T.; Julian, R. R., Using ESI-MS to Probe Protein Structure by Site-Specific Noncovalent Attachment of 18-Crown-6. *Journal of the American Society for Mass Spectrometry* **2006,** *17* (9), 1209-1215.

11. Mendoza, V. L.; Vachet, R. W., Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrometry Reviews* **2009,** *28* (5), 785-815.

12. Hanai, R.; Wang, J. C., Protein footprinting by the combined use of reversible and irreversible lysine modifications. *Proceedings of the National Academy of Sciences of the United States of America* **1994,** *91* (25), 11904-11908.

13. Sharp, J. S.; Becker, J. M.; Hettich, R. L., Analysis of Protein Solvent Accessible Surfaces by Photochemical Oxidation and Mass Spectrometry. *Analytical Chemistry* **2004,** *76* (3), 672-683.

14. Shi, Y.; Fernandez-Martinez, J.; Tjioe, E.; Pellarin, R.; Kim, S. J.; Williams, R.; Schneidman-Duhovny, D.; Sali, A.; Rout, M. P.; Chait, B. T., Structural Characterization by Cross-linking Reveals the Detailed Architecture of a Coatomer-related Heptameric Module from the Nuclear Pore Complex. *Molecular & Cellular Proteomics* **2014,** *13* (11), 2927-2943.

15. Pacheco, B.; Maccarana, M.; Goodlett, D. R.; Malmström, A.; Malmström, L., Identification of the Active Site of DS-epimerase 1 and Requirement of N-Glycosylation for Enzyme Function. *Journal of Biological Chemistry* **2009,** *284* (3), 1741-1747.

16. Kaur, P.; Kiselar, J.; Yang, S.; Chance, M. R., Quantitative Protein Topography Analysis and High-Resolution Structure Prediction Using Hydroxyl Radical Labeling and Tandem-Ion Mass Spectrometry (MS). *Molecular & Cellular Proteomics* **2015,** *14* (4), 1159-1168.

17. Hambly, D.; Gross, M., Laser flash photochemical oxidation to locate heme binding and conformational changes in myoglobin. *International Journal of Mass Spectrometry* **2007,** *259* (1), 124-129.

18. Guan, J.-Q.; Vorobiev, S.; Almo, S. C.; Chance, M. R., Mapping the G-Actin Binding Surface of Cofilin Using Synchrotron Protein Footprinting. *Biochemistry* **2002,** *41* (18), 5765-5775.

19. Chen, Z. A.; Pellarin, R.; Fischer, L.; Sali, A.; Nilges, M.; Barlow, P. N.; Rappsilber, J., Structure of Complement C3(H2O) Revealed By Quantitative Cross-Linking/Mass Spectrometry And Modeling. *Molecular & Cellular Proteomics* **2016,** *15* (8), 2730-2743.

20. Jones, L. M.; B. Sperry, J.; A. Carroll, J.; Gross, M. L., Fast Photochemical Oxidation of Proteins for Epitope Mapping. *Analytical Chemistry* **2011,** *83* (20), 7657-7661.

21. Sheshberadaran, H.; Payne, L. G., Protein antigen-monoclonal antibody contact sites investigated by limited proteolysis of monoclonal antibody-bound antigen: protein "footprinting". *Proceedings of the National Academy of Sciences of the United States of America* **1988,** *85* (1), 1-5.

22. Steiner, R. F.; Albaugh, S.; Fenselau, C.; Murphy, C.; Vestling, M., A mass spectrometry method for mapping the interface topography of interacting proteins, illustrated by the melittin-calmodulin system. *Analytical Biochemistry* **1991,** *196* (1), 120-125.

23. Wang, L.; Chance, M. R., Structural Mass Spectrometry of Proteins Using Hydroxyl Radical Based Protein Footprinting. *Analytical Chemistry* **2011,** *83* (19), 7234-7241.

24. Xu, G.; Chance, M. R., Radiolytic Modification and Reactivity of Amino Acid Residues Serving as Structural Probes for Protein Footprinting. *Analytical Chemistry* **2005,** *77* (14), 4549-4555.

25. Maleknia, S. D.; Brenowitz, M.; Chance, M. R., Millisecond radiolytic modification of peptides by synchrotron X-rays identified by mass spectrometry. *Analytical Chemistry* **1999,** *71* (18), 3965-3973.

26. Hambly, D. M.; Gross, M. L., Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale. *Journal of the American Society for Mass Spectrometry* **2005,** *16* (12), 2057-2063.

27. Li, K. S.; Shi, L.; Gross, M. L., Mass Spectrometry-Based Fast Photochemical Oxidation of Proteins (FPOP) for Higher Order Structure Characterization. *Accounts of Chemical Research* **2018**.

28. Alexander, N. S.; Stein, R. A.; Koteiche, H. A.; Kaufmann, K. W.; McHaourab, H. S.; Meiler, J., RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics. *PLOS ONE* **2013,** *8* (9), e72851.

29. Fischer, A. W.; Alexander, N. S.; Woetzel, N.; Karakas, M.; Weiner, B. E.; Meiler, J., BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints. *Proteins* **2015,** *83* (11), 1947-1962.

30. 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. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** *105* (12), 4685-4690.

31. Sgourakis, N. G.; Lange, O. F.; DiMaio, F.; André, I.; Fitzkee, N. C.; Rossi, P.; Montelione, G. T.; Bax, A.; Baker, D., Determination of the Structures of Symmetric Protein Oligomers from NMR Chemical Shifts and Residual Dipolar Couplings. *Journal of the American Chemical Society* **2011,** *133* (16), 6288-6298.

32. Huang, W.; Ravikumar, K. M.; Parisien, M.; Yang, S., Theoretical modeling of multiprotein complexes by iSPOT: Integration of small-angle X-ray scattering, hydroxyl radical footprinting, and computational docking. *Journal of Structural Biology* **2016,** *196* (3), 340-349.

33. Rossi, P.; Shi, L.; Liu, G.; Barbieri, C. M.; Lee, H.-W.; Grant, T. D.; Luft, J. R.; Xiao, R.; Acton, T. B.; Snell, E. H.; Montelione, G. T.; Baker, D.; Lange, O. F.; Sgourakis, N. G., A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins: Structure, Function, and Bioinformatics* **2015,** *83* (2), 309-317.

34. Putnam, D. K.; Weiner, B. E.; Woetzel, N.; Lowe, E. W.; Meiler, J., BCL::SAXS: GPU accelerated Debye method for computation of small angle X-ray scattering profiles. *Proteins* **2015,** *83* (8), 1500-1512.

35. Schneidman-Duhovny, D.; Kim, S. J.; Sali, A., Integrative structural modeling with small angle X-ray scattering profiles. *BMC Structural Biology* **2012,** *12*, 17.

36. DiMaio, F.; Tyka, M. D.; Baker, M. L.; Chiu, W.; Baker, D., Refinement of protein structures into low-resolution density maps using rosetta. *Journal of Molecular Biology* **2009,** *392* (1), 181-190.

37. Leelananda, S. P.; Lindert, S., Iterative Molecular Dynamics–Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5131-5145.

38. Lindert, S.; McCammon, J. A., Improved cryoEM-Guided Iterative Molecular Dynamics--Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction. *Journal of Chemical Theory and Computation* **2015,** *11* (3), 1337-1346.

39. Lindert, S.; Alexander, N.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J., EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure (London, England: 1993)* **2012,** *20* (3), 464-478.

40. 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* (9), 669-677.

41. DiMaio, F.; Song, Y.; Li, X.; Brunner, M. J.; Xu, C.; Conticello, V.; Egelman, E.; Marlovits, T.; Cheng, Y.; Baker, D., Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nature Methods* **2015,** *12* (4), 361-365.

42. Jiang, W.; Baker, M. L.; Ludtke, S. J.; Chiu, W., Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *Journal of Molecular Biology* **2001,** *308* (5), 1033-1044.

43. Baker, M. L.; Ju, T.; Chiu, W., Identification of secondary structure elements in intermediate-resolution density maps. *Structure (London, England: 1993)* **2007,** *15* (1), 7-19.

44. Kahraman, A.; Herzog, F.; Leitner, A.; Rosenberger, G.; Aebersold, R.; Malmström, L., Cross-Link Guided Molecular Modeling with ROSETTA. *PLOS ONE* **2013,** *8* (9), e73411.

45. Kahraman, A.; Malmström, L.; Aebersold, R., Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics* **2011,** *27* (15), 2163-2164.

46. Walzthoeni, T.; Joachimiak, L. A.; Rosenberger, G.; Röst, H. L.; Malmström, L.; Leitner, A.; Frydman, J.; Aebersold, R., xTract: software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nature Methods* **2015,** *12* (12), 1185.

47. Herzog, F.; Kahraman, A.; Boehringer, D.; Mak, R.; Bracher, A.; Walzthoeni, T.; Leitner, A.; Beck, M.; Hartl, F.-U.; Ban, N.; Malmström, L.; Aebersold, R., Structural Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science* **2012,** *337* (6100), 1348-1352.

48. Webb, B.; Viswanath, S.; Bonomi, M.; Pellarin, R.; Greenberg, C. H.; Saltzberg, D.; Sali, A., Integrative structure modeling with the Integrative Modeling Platform. *Protein Science*, n/a-n/a.

49. Politis, A.; Schmidt, C.; Tjioe, E.; Sandercock, A. M.; Lasker, K.; Gordiyenko, Y.; Russel, D.; Sali, A.; Robinson, C. V., Topological models of heteromeric protein assemblies from mass spectrometry: application to the yeast eIF3:eIF5 complex. *Chemistry & Biology* **2015,** *22* (1), 117-128.

50. Saltzberg, D. J.; Broughton, H. B.; Pellarin, R.; Chalmers, M. J.; Espada, A.; Dodge, J. A.; Pascal, B. D.; Griffin, P. R.; Humblet, C.; Sali, A., A Residue-Resolved Bayesian Approach to Quantitative Interpretation of Hydrogen–Deuterium Exchange from Mass Spectrometry: Application to Characterizing Protein–Ligand Interactions. *The Journal of Physical Chemistry B* **2017,** *121* (15), 3493-3501.

51. Zeng-Elmore, X.; Gao, X.-Z.; Pellarin, R.; Schneidman-Duhovny, D.; Zhang, X.-J.; Kozacka, K. A.; Tang, Y.; Sali, A.; Chalkley, R. J.; Cote, R. H.; Chu, F., Molecular Architecture of Photoreceptor Phosphodiesterase Elucidated by Chemical Cross-Linking and Integrative Modeling. *Journal of Molecular Biology* **2014,** *426* (22), 3713-3728.

52. Webb, B.; Lasker, K.; Velázquez-Muriel, J.; Schneidman-Duhovny, D.; Pellarin, R.; Bonomi, M.; Greenberg, C.; Raveh, B.; Tjioe, E.; Russel, D.; Sali, A., Modeling of proteins and their assemblies with the Integrative Modeling Platform. *Methods in Molecular Biology (Clifton, N.J.)* **2014,** *1091*, 277-295.

53. Xie, B.; Sood, A.; Woods, R. J.; Sharp, J. S., Quantitative Protein Topography Measurements by High Resolution Hydroxyl Radical Protein Footprinting Enable Accurate Molecular Model Selection. *Scientific Reports* **2017,** *7* (1), 4552.

54. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., Chapter nineteen - Rosetta3:

An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In *Methods in Enzymology*, Brand, M. L. J. a. L., Ed. Academic Press: 2011; Vol. 487, pp 545-574.

55. Huang, W.; Ravikumar, Krishnakumar M.; Chance, Mark R.; Yang, S., Quantitative Mapping of Protein Structure by Hydroxyl Radical Footprinting-Mediated Structural Mass Spectrometry: A Protection Factor Analysis. *Biophysical Journal* **2015,** *108* (1), 107-115.

56. Gustavsson, M.; Wang, L.; Gils, N. v.; Stephens, B. S.; Zhang, P.; Schall, T. J.; Yang, S.; Abagyan, R.; Chance, M. R.; Kufareva, I.; Handel, T. M., Structural basis of ligand interaction with atypical chemokine receptor 3. *Nature Communications* **2017,** *8*, 14135.

57. Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D., Protein Structure Prediction Using Rosetta. Enzymology, B. T. M. i., Ed. Academic Press: 2004; Vol. 383, pp 66-93.

58. 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 functions11Edited by F. E. Cohen. *Journal of Molecular Biology* **1997,** *268* (1), 209-225.

59. Bradley, P.; Misura, K. M. S.; Baker, D., Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science* **2005,** *309* (5742), 1868-1871.

60. Bender, B. J.; Cisneros, A.; 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* (34), 4748-4763.

61. Kim, D. E.; Chivian, D.; Baker, D., Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* **2004,** *32* (suppl_2), W526-W531.

62. 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.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017,** *13* (6), 3031-3048.

63. Tyka, M. D.; Keedy, D. A.; André, I.; Dimaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D., Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology* **2011,** *405* (2), 607-618.

64. Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D., Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science* **2014,** *23* (1), 47-55.

65. Durham, E.; Dorr, B.; Woetzel, N.; Staritzbichler, R.; Meiler, J., Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of Molecular Modeling* **2009,** *15* (9), 1093-1108.

66. Street, A. G.; Mayo, S. L., Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design* **1998,** *3* (4), 253-258.

67. Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **2017,** *357* (6347), 168-175.

68. London, N.; Schueler-Furman, O., Funnel Hunting in a Rough Terrain: Learning and Discriminating Native Energy Funnels. *Structure* **2008,** *16* (2), 269-279.

69. 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., 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* (7625), 329-335.

70. Plaxco, K. W.; Simons, K. T.; Baker, D., Contact order, transition state placement and the refolding rates of single domain proteins11Edited by P. E. Wright. *Journal of Molecular Biology* **1998,** *277* (4), 985-994.

71. Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. Columbus OH: Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73.

FOR TABLE OF CONTENTS USE ONLY

Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data

Melanie L. Aprahamian, Emily E. Chea, Lisa Jones, Steffen Lindert