ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Article

Extracting Residue Solvent Exposure from Covalent Labeling Data with Machine Learning: A Hybrid Approach for Protein Structure Prediction

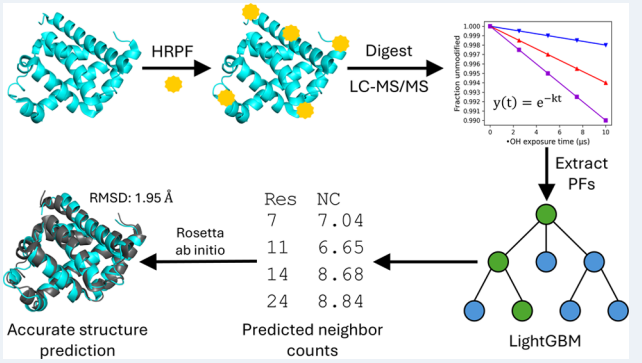
Elijah H. Day and Steffen Lindert*

Cite This: J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Hydroxyl radical protein footprinting (HRPF) coupled with mass spectrometry yields information about residue solvent exposure and protein topology. However, data from these experiments are sparse and require computational interpretation to generate useful structural insight. We previously implemented a Rosetta algorithm that uses experimental HRPF data to improve protein structure prediction. Modern structure prediction methods, such as AlphaFold2 (AF2), use machine learning (ML) to generate their predictions. Implementation of an HRPF-guided version of AF2 is challenging due to the substantial amount of training data required and the inherently abstract nature of ML networks. Thus, here we present a hybrid method that uses a light gradient boosting machine to predict residue solvent accessibility from experimental HRPF data. These predictions were subsequently used to improve Rosetta structure prediction. Our hybrid approach identified models with atomic-level detail for all four proteins in our benchmark set. These results illustrate that it is possible to successfully use ML in combination with HRPF data to accurately predict protein structures.



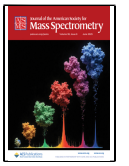
INTRODUCTION

Proteins play an essential role in nearly every cellular process, making a detailed understanding of their structures critical for unraveling complex biological pathways and advancing drug development. Classical structural biology techniques such as nuclear magnetic resonance (NMR), cryo-electron microscopy (cryo-EM), and X-ray crystallography are routinely used to elucidate protein structures with atomic-level resolution.[1−3] However, each of these methods comes with its own set of challenges, such as size constraints or complex sample preparation, which can severely limit their applicability. As an alternative to experimental structure determination, modern computational methods have enabled the prediction of protein structures with nearly atomic-level accuracy. AlphaFold2 (AF2), a neural network developed by DeepMind, is able to predict structures with a level of accuracy that rivals that of experimental methods.[4,5] However, there are cases in which AF2 still struggles to predict accurate structures, primarily membrane proteins, large protein complexes, and proteins with distinct conformational states.[6]

Mass spectrometry (MS) is a versatile analytical tool that has become invaluable in the field of structural biology.[7] Several techniques, collectively known as structural MS, have been developed to probe different structural characteristics of proteins and their complexes.[8] One key advantage of these techniques is that they can be applied to proteins that are traditionally challenging to investigate using classical techniques, such as membrane proteins and intrinsically disordered proteins (IDPs).[9,10] Additionally, many structural MS techniques do not require a large amount of sample and can be used on heterogeneous protein mixtures. The most widely used structural MS techniques include chemical cross-linking-MS (XL-MS),[11] hydrogen−deuterium exchange-MS (HDX-MS),[12,13] ion mobility-MS (IM-MS),[14] surface-induced dissociation-MS (SID-MS),[15] and covalent labeling-MS (CL-MS).[16,17]

Covalent labeling is a powerful technique in which proteins in solution are exposed to highly reactive labeling reagents that irreversibly modify amino acid side chains. Residues that are more solvent-exposed are typically labeled at a higher rate, yielding valuable insight into both local residue environment and overall protein topology. Labeling can be achieved using a variety of both specific and nonspecific reagents. Commonly used reagents include hydroxyl radicals, trifluoromethyl radicals, carbenes, and diethylpyrocarbonate (DEPC).[18−22] Among

Received: February 18, 2025
Revised: April 18, 2025
Accepted: May 8, 2025
Published: May 20, 2025



Downloaded via UNIV OF CALIFORNIA LOS ANGELES on August 14, 2025 at 16:09:23 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications

© 2025 American Society for Mass Spectrometry. Published by American Chemical Society. All rights reserved.

these, hydroxyl radical protein footprinting (HRPF) has emerged as a particularly robust and widely used technique.[23] There are numerous approaches for generating hydroxyl radicals, with the most common being the X-ray photolysis of water and the laser-based fast photochemical oxidation of proteins (FPOP).[24,25] These radicals can theoretically oxidize 19 of the 20 naturally occurring amino acids with varying degrees of reactivity, although oxidation products are readily detected for only 14 residues.[26] The small size of hydroxyl radicals, similar to that of water, allows them to access all solvent-accessible areas of a protein.[27,28] This, along with their high degree of reactivity, makes hydroxyl radicals an ideal probe for CL-MS experiments.[29,30]

Despite the promising utility of structural MS, the data collected from these techniques cannot alone be used to unambiguously determine tertiary structures.[31,32] Thus, computational interpretation is needed to generate useful structural insights from these data. Previous work has shown that structural information extracted from sparse MS data can be used to improve prediction accuracy across a variety of data types and modeling platforms.[33−53] In our previous work, we demonstrated that the correlation between residue solvent accessibility and experimental HRPF protection factor (PF) data can be leveraged to create a Rosetta scoring term to improve ab initio structure prediction.[46] We have also shown that the magnitude of this improvement can be increased by accounting for side-chain flexibility through the use of Rosetta mover models.[54]

While machine learning (ML)-based approaches, such as AF2, can predict protein structures with remarkable accuracy, it remains imperative to explore how sparse experimental data can complement these methods and further refine the predictions. While ML tools that leverage the power of experimental data, such as AlphaLink (a version of AF2 that incorporates experimental XL-MS data), have been developed, such networks typically require large, specialized data sets for training and careful treatment of the experimental data.[55−58] The direct integration of raw experimental data into ML prediction networks is impossible; instead, the data must undergo a transformation and embedding process to align with the model architecture while preserving critical structural information. This is particularly challenging for AF2-based prediction networks, as their predictions rely heavily on coevolutionary information extracted from multiple sequence alignments (MSAs). In contrast, HRPF experiments yield information about residue exposure, which fundamentally differs from sequence alignment-based inputs. Thus, transforming such data into a format that integrates seamlessly with sequence-derived inputs requires significant effort. This limitation led us to investigate whether a hybrid approach, combining the strengths of machine learning with the versatility of Rosetta, could offer a more accessible alternative for improving structure prediction, particularly in cases where training data are sparse or not easily representable within the existing model architecture.

In this work, we have probed whether a gradient boosting regression tree model that predicts residue solvent accessibility using HRPF labeling data could be used to improve Rosetta ab initio structure prediction. Our pipeline uses experimental labeling data, as well as additional features extracted from the query sequence to generate expected residue neighbor counts that are then used in a Rosetta score term that evaluates decoys based on their agreement to the expected value. We benchmarked our score term on a set of four proteins with experimental HRPF data available in the literature. The inclusion of our score term allowed us to identify models with low (<5 Å) root-mean-square deviation (RMSD) for all four benchmark proteins, compared to only one when HRPF data are not used. Our hybrid approach yielded a higher average RMSD improvement across our benchmark set than our previous HRPF-guided prediction methods, without the need for Rosetta mover models, suggesting that elements of machine learning can improve non-machine learning-based prediction methods. The *hrf_GBM_rescore* algorithm is freely available through the Rosetta software suite, and the regression model is available on GitHub (https://github.com/LindertLab/HRF_GBM_Rescore). In addition, an in-depth tutorial describing the modeling protocol is available in Supplementary Note 1.

METHODS

Labeling and Exposure Metrics. We used PF, a metric originally introduced by Huang and co-workers, in this work.[59] PF allows for the direct comparison of HRPF data between sites within a protein by relating the footprinting rate constant of a residue to its experimentally determined intrinsic reactivity. We employed the natural logarithm of PF (eq 1) for structure prediction, as it has been shown to have a moderate correlation with residue solvent exposure.

$$\ln(\mathrm{PF}) = \ln \frac{R_i}{k_i} \tag{1}$$

R_i is the intrinsic reactivity of residue i toward labeling, and k_i is the experimentally determined footprinting rate constant. As FPOP experiments measure residue modification at a single time point, footprinting rate constants cannot be directly determined.[25] However, it has been demonstrated that the raw modification extent (%) from FPOP experiments can be used instead of rate constants for PF calculations.

Conical neighbor count, a metric introduced in our previous work, was used in this study due to its correlation with ln(PF) and its low computational cost compared to explicit solvent-accessible surface area (SASA) calculations.[31] Conceptually, the neighbor count serves as a proxy for residue solvent exposure: residues buried within the protein core tend to have more neighboring residues, while those nearer to the solvent interface have fewer. While traditional SASA calculations require the explicit modeling of interactions between the protein surface and a solvent probe, neighbor count estimates residue accessibility using pairwise geometric relationships. The conical version refines this approach by only counting neighbors located in the direction of the side chain, using a cone aligned with the $C\alpha-C\beta$ vector. This avoids inflating neighbor counts for residues that face outward but have atoms positioned behind them. Conical neighbor count for a residue i is defined as the product of two sigmoidal functions, one angle-based and one distance-based, summed over all other residues. The angle contribution $A(\Theta_{ij})$ is shown in eq 2.

$$A(\Theta_{ij}) = \frac{1}{1 + \exp(S \times (\Theta_{ij} - M))} \tag{2}$$

Here, S represents the angle steepness value, and M represents the angle midpoint values. We used an angle steepness value of 2π and an angle midpoint value of $\pi/2$ for neighbor count calculations. These values were used in our previous work and were found to maximize the correlation between residue neighbor count and PF.[54] Calculations of conical neighbor

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

counts were done by using the Rosetta *per_residue_solvent_exposure* application.

Development and Evaluation of a Gradient Boosting Regression Tree Model for Neighbor Count Prediction. To explore whether machine learning methods could improve predictions of labeled residue neighbor counts from experimental HRPF data, we trained a light gradient boosting machine (LightGBM) regression model. LightGBM is an efficient implementation of gradient boosting, a method that builds an ensemble of weak learners, typically decision trees, to construct a robust prediction model.[60] In gradient boosting, each successive tree is trained to correct the residual errors of its predecessors by iteratively minimizing an arbitrary loss function, which ensures that the model progressively improves throughout training.[61] We chose to minimize the root-mean-square error (RMSE) between the predicted and observed neighbor counts during training, as we wanted to more heavily penalize outliers.

Curation of a LightGBM Training Set. We conducted an extensive literature search to identify a data set of HRPF-labeled proteins suitable for training our model. We required proteins in our data set to have high-resolution structures available in the PDB, as to allow for accurate neighbor count calculations for the labeled residues. The data set included a combination of FPOP and synchrotron labeling data, ensuring that our model is robust across different experimental conditions and labeling techniques. The full synchrotron and FPOP data sets used for model training can be found in Tables S1 and S2, respectively.

Labeling data for CXCL12 (PDB ID: 1A15; 10 labeled residues) were extracted from the manuscript of Gustavsson et al.[62] Barstar (PDB ID: 1A19; 21 labeled residues) data were gathered from the work of Gau and co-workers.[63] PFs for both active (PDB ID: 1AVP; 6 labeled residues) and inactive (PDB ID: 4EKF; 7 labeled residues) human AVP were extracted from the work of Gupta et al.[64] Labeling data for α1-antitrypsin (PDB ID: 1QLP; 11 labeled residues) were extracted from the manuscript of Zheng et al.[65] PFs for POT1 (PDB ID: 1XJV; 39 labeled residues) were calculated from the rate constants reported in the work of Xu and co-workers.[66] Cytochrome C data (PDB ID: 2B4Z; 26 labeled residues) were extracted from the manuscript of Chea et al.[67] Residue-level PF data for the kinase (PDB ID: 2V5Q; 40 labeled residues) and polo-box domains (PDB ID: 3HIK; 25 labeled residues) of PLK 1 were gathered from the work of Ruan et al.[68] Labeling data for both WT (PDB ID: 2XKS; 17 labeled residues) and ΔN6 (PDB ID: 2XKU; 16 labeled residues) β2-microglobulin were collected from the manuscript of Cornwell and co-workers.[69] PF data for CheY (PDB ID: 3CHY; 19 labeled residues) were extracted from the work of Hamid et al.[70] Gelsolin (PDB ID: 3FFN; 8 labeled residues) PF data were collected from the work of Kiselar and colleagues.[71]

Generation of Initial Sequence Descriptors and Template Features. In addition to the experimental HRPF data, we explored various sequence and template-based features that have been used in other solvent-accessibility prediction models. For each labeled residue in our training set, we generated an initial feature set of 423 descriptors. These descriptors included experimentally determined PF, residue identity, physicochemical properties, position-specific scoring matrix (PSSM) profiles, predicted secondary structure, predicted disorder, and the homologous side-chain environment (SCE). Residue identity and physicochemical properties were encoded into a 12-dimensional feature vector, incorporating attributes such as residue mass, hydrophobicity, intrinsic reactivity toward hydroxyl radicals, and polarizability. Additionally, we included free-residue solvent-accessible surface area, dipole moment, pI, pK_a, pK_b, and secondary structure propensities among the encoded features.

PSSM profiles were generated using PSI-BLAST and contain information regarding the conservation of amino acids at a given position within a generated multiple sequence alignment.[72,73] The scores within the PSSM represent how frequently each residue appears at a particular position relative to the baseline frequency.[74,75] PSSMs were generated by running PSI-BLAST on the query sequence for three iterations with an *E*-value cutoff of 0.001 against the nonredundant NCBI data set. Substitution scores extracted from the PSSM profiles were normalized via a standard logistic function (eq 3).

$$\mathrm{PSSM}' = \frac{1}{1 + \exp(-\mathrm{PSSM})} \tag{3}$$

Here, PSSM is the score extracted from the PSSM itself, and PSSM′ is the normalized value. A total of 41 features were extracted from the PSSM for each residue: 20 normalized substitution scores, 20 absolute substitution counts, and an overall information per position score that was not normalized prior to encoding.

Predicted secondary structure assignments for the training set were generated using the PSIPRED program.[76] For each residue, PSIPRED returns three probability scores, each representing the likelihood of the residue adopting a coil, helix, or strand structure. These probabilities were encoded into a three-dimensional feature vector. Residue disorder was predicted using the DISOPRED program.[77] Disordered regions tend to be more solvent-exposed compared to structured regions, which makes disorder a potentially informative feature of our model. DISOPRED assigns a score between 0 and 1 to each residue, with lower scores indicating higher levels of disorder.

Side-chain environment (SCE) is a descriptor first introduced by Bowie et al. to identify and categorize sequences that fold into known tertiary structures.[78] Residue SCE has successfully been used in numerous modeling tasks and binding-site prediction models.[79,80] Residues are broadly categorized as either exposed, partially buried, or buried based on their relative solvent-accessible surface areas. The partially buried and buried environments are then further subdivided according to the fraction of the side-chain area covered by polar atoms for a total of six possible SCE assignments.

For each residue in our training set, we encoded a template-based feature that we have called "homologous SCE". Template structures were found by performing a BLASTP against the NCBI data set of PDB structures.[81] The 10 hits with the lowest *E*-values (excluding the query structure) were downloaded for processing and SCE calculation. If less than 10 hits were available, all hits were used for SCE calculations. The sequence for each hit was aligned with the query sequence, and all residues that did not have a perfect identity match were filtered out. SCEs were assigned for template residues based on SASAs and polar fractions calculated with the program NACCESS.[82] For a labeled residue *i*, SCE was assigned based on the most returned SCE assignment across all templates.

We opted to use a sliding-window approach for predicting the neighbor count of HRPF-labeled residues, as it has been theorized that residue microenvironment and sequence context can influence residue labeling yield.[83] Instead of solely relying on

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

the features extracted for a labeled residue *i*, we also encoded features from the adjacent residues on either side of *i*, resulting in a local-sequence fragment of input features. Local window sizes ranging from 3 to 15 residues were evaluated to optimize predictive performance. A window size of 7 residues was found to maximize the predictive performance of the model.

Additional sequence embeddings were extracted using the R package "Peptides".[84] For each local-sequence window, we extracted Cruciani properties, FASGAI vectors, Kidera factors, Z-scales, VSHE-scales, T-scales, ST-scales, protFP descriptors, and BLOSUM indices.[85−93] A total of 66 features were extracted for each peptide window.

LightGBM Feature Selection and Training. We initially trained an ensemble of LightGBM models using all 423 features and a 10-fold cross-validation strategy. In 10-fold cross-validation, the training data set is randomly split into 10 equal-sized folds. A model is then trained using 9 of the folds as training data, and the resulting model is tested on the held-out fold. This process is repeated until each fold has served as the test set. For each model, we computed feature importance metrics reflecting the contribution of each feature to model predictions. This process was repeated 100 times with different fold splits, generating a total of 1000 models. By averaging the feature importance scores across all models, we created an initial ranking of the 423 features.

Using this ranking, we aimed to identify the optimal feature set. Starting with the most important feature, we performed successive rounds of 10-fold cross-validation, adding one feature to the training set at a time. After each round, we evaluated the mean cross-validation score (RMSE). This process was repeated 25 times, and the point at which the average cross-validation score reached a minimum was identified as the optimal number of features. Cross-validation score versus number of features plots for all 25 trials can be found in Figure S1.

Given the small size of our training data set, we employed a leave-one-out cross-validation (LOOCV) strategy for hyperparameter tuning. LOOCV is a stringent form of cross-validation where a separate model is trained and evaluated for each labeled residue in the training set.[94] In each iteration, one sample is held out as the test set, while the model is trained on the remaining data. The model is then evaluated on the held-out sample. This process is repeated for every sample, and the average of the evaluation metrics across all iterations is reported as the final performance measure. We combined LOOCV with a grid search over possible hyperparameter values to identify the hyperparameter set that minimized the LOOCV score. The final regression model was trained using these hyperparameters and the full training data set. The most important features in the final LightGBM model included the predicted SCE, experimentally determined PF, and predicted secondary structure. Several PSSM-derived features also contributed significantly to the model performance. Complete feature importance rankings for the final model can be found in Figure S2.

Structure Prediction Benchmark Set. Our benchmark set for structure prediction was composed of four well structured, globular proteins that have residue-resolved HRPF data available in the literature. We required our benchmark set to contain HRPF data gathered via a variety of experimental methods (FPOP and synchrotron radiolysis), as we wanted to ensure that our method is broadly applicable. PFs for myoglobin (PDB ID: 1YMB) and lysozyme (PDB ID: 2LYZ) were gathered from the work of Xie and co-workers.[47] We extracted PF data for 43 residues in myoglobin and 34 residues in lysozyme. PF data for the Ca^{2+}-bound compact form of calmodulin (PDB ID: 1PRW) were obtained from the work of Kaur et al.; PFs for 22 labeled residues were extracted.[95] Labeling data for LMPTP (PDB ID: 5JNS) were gathered from the manuscript of Stanford et al.; PFs for 25 residues were obtained.[96] These proteins were previously used to benchmark our HRPF-guided structure prediction methods, making them ideal candidates for benchmarking our new approach, as they allow for direct performance comparisons between the two methods.[46,54] The HRPF data used for the benchmark set can be found in Tables S3 and S4.

Generation of Ab Initio Models. The Rosetta *AbInitio Relax* protocol was used to generate 20,000 decoys for each protein in the benchmark set.[97] The use of 20,000 decoys per target ensures thorough sampling and enhances the reliability of downstream model comparisons. Native FASTA sequences were obtained from their respective PDB entries and were used to create the fragment libraries required for decoy assembly.[98] Fragment libraries were generated using the Robetta web server.[99] Decoys were generated by using the default ab initio parameters, followed by relaxation and scoring with the default Ref15 score function. The crystal structure of each benchmark protein was supplied and used only for the RMSD calculations. No experimental data were used during fragment generation and decoy assembly.

Determination of Residue Types Used for Rescoring. Nineteen of the 20 naturally occurring amino acid residues can theoretically be modified by hydroxyl radicals with varying levels of reactivity. However, oxidation products are readily detected for only 14 residues.[26] Thus, we sought to find a subset of residues that would be most useful in structure prediction tasks. We eliminated residues with an intrinsic reactivity (R_i) lower than that of proline from our consideration, as these oxidation products are rarely observed and do not correlate well with residue structural features.[47] We also excluded serine and threonine residues, as they have been reported to have unclear oxidation products, which may complicate data interpretation.[23] Cysteine and methionine residues were also eliminated, as they have such a high reactivity toward hydroxyl radicals that they often are modified even while completely buried.[100] Thus, we investigated a set of 10 residues with sequential reactivities: W, Y, F, H, L, I, R, K, V, and P. We examined the performance of the following residue type combinations: WYF, WYFH, WYFHL, WYFHLI, WYFHLIR, WYFHLIRK, WYFHLIRKV, and WYFHLIRKVP. For each residue combination, we predicted conical neighbor counts using our gradient boosting model for each labeled residue in the benchmark set. Observed neighbor counts were calculated using the benchmark set crystal structures. Residue sets were evaluated based on the RMSE and R^2 values calculated between the predicted and observed neighbor counts for all labeled residues within the residue set.

Rescoring of Ab Initio Models. We implemented a new scoring term, *hrf_GBM*, to assess the agreement between predicted and observed decoy residue neighbor counts. For each labeled residue, predicted neighbor counts were generated by using a gradient boosting regression model. Observed neighbor counts were calculated from the decoy structure using the Rosetta *per_residue_solvent_exposure* application. The absolute difference between the observed and predicted neighbor counts $|diff_i|$ is calculated for each labeled residue *i*. Based on these differences, the total *hrf_GBM* score for a structure is defined as the sum of all per-residue contributions, as defined in eq 4:

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

$$hrf_GBM = \sum_{i}^{n} \frac{-1.0}{1.0 + \exp(2.9(|diff_i| - 2.1))} \quad (4)$$

where n is the number of labeled residues, 2.9 is the steepness value, and 2.1 is the delta value. The score term is designed such that each residue will be assigned a score between 0, indicating complete disagreement with the experimental data, and −1, indicating perfect agreement with the experimental data.

The *hrf_GBM* score term was used to rescore sets of 20,000 ab initio models for each benchmark protein. The total score for a given decoy was calculated as a weighted sum of the new score term and the Rosetta Ref15 score, as shown in eq 5:

$$\text{total score} = (9 \times hrf_GBM) + \text{Ref15} \quad (5)$$

Data Availability. In addition to the Supporting Information, all experimental HRPF data used for LightGBM training can be found in refs 62−71 and refs 47, 95, and 97 for the structure prediction benchmark set.

Code Availability. The *hrf_GBM* score term is available for free academic use through the Rosetta software suite at https://www.rosettacommons.org/software/. The pretrained LightGBM model is available at https://github.com/LindertLab/HRF_GBM_Rescore.

RESULTS AND DISCUSSION

In our previous work, we demonstrated that the approximation of side-chain dynamics via the generation of Rosetta mover models could be used to improve the correlation between residue neighbor count and experimental HRPF data.[54] We found that this correlation was particularly strong for residue types W, Y, F, H, and L. This relationship was leveraged to construct a scoring term that rewarded residues based on their agreement to the experimental HRPF data. While our previous approach was effective in improving prediction accuracy, it was limited to a small subset of highly reactive residues and required the generation of mover models, which is computationally expensive and requires proficiency with Rosetta. Additionally, it has been shown that residue solvent accessibility alone is insufficient for fully explaining the labeling extent of a given residue.[67,83] Thus, we hypothesized that if we were able to more accurately predict the neighbor count of labeled residues using HRPF data, we would both improve our predictive powers and broaden the subset of residues from which we can extract useful structural information from.

Development of a Gradient Boosting Regression Model Improved Prediction of Labeled Residue Conical Neighbor Count. We hypothesized that a machine learning regression model could be used to better predict the conical neighbor count of HRPF-labeled residues, thereby improving the overall predictive power of HRPF data. Our previous work has shown that there is a moderate correlation between the natural log of the PF and the conical neighbor count of labeled residues; however, this correlation was limited to certain residue types.[54] Additionally, it has been reported that the labeling rates of residues toward other CL-MS reagents can be highly dependent on local environment and sequence context.[39,83] This prompted us to explore whether machine learning could better capture these complex relationships within HRPF data.

To explore our proposed hypothesis, we trained a LightGBM model to predict residue neighbor counts using experimental PF data, alongside other sequence and template-based features. We trained our model on a data set of proteins with residue-resolved HRPF data available in the literature. In contrast to our previous work, which only investigated a subset of highly reactive residues (WYFHL), we trained our model on all HRPF-labeled residues regardless of side-chain identity. We then used the model to predict labeled residue neighbor counts for a benchmark set of four proteins not used for LightGBM training: myoglobin (PDB ID: 1YMB), calmodulin (Ca^{2+}-bound, compact form; PDB ID: 1PRW), lysozyme (PDB ID: 2LYZ), and low molecular weight protein tyrosine phosphatase (LMPTP; PDB ID: 5JNS). These proteins have residue-resolved HRPF data available in the literature, each with at least 15 labeled residues. Furthermore, these four benchmark proteins were used in our previous evaluations of HRPF-guided structure prediction methods, allowing for a direct comparison between the performance of our new model and earlier approaches.[46,54]

Calmodulin, in particular, is an ideal benchmark protein due to its conformational flexibility. In the presence of calcium, calmodulin samples a wide range of conformations due to its highly flexible linker region. The compact, Ca^{2+}-bound structure used in our benchmark set has been shown to represent a prominent solution state found in calcium-saturated conditions.[101−103] As such, it provides a valuable test case to probe the ability of integrative modeling approaches to guide sampling toward biologically relevant conformations.

Our predictions for labeled residue neighbor counts yielded a low RMSE (1.80) and high R^2 value (0.60) when compared to the crystal structure neighbor counts (Figure 1A). We observed that residues with a higher intrinsic reactivity toward hydroxyl radicals generally exhibited lower prediction RMSEs compared to residues with a lower reactivity. Figure 2A−C shows the prediction RMSEs for residues with high, medium, and low reactivities toward hydroxyl radicals. Highly reactive residues exhibited a prediction RMSE of 1.6, while medium and low reactivity residues yielded prediction RMSEs of 1.65 and 2.02, respectively. We performed a systematic search of residue type combinations and found that the group containing residue types W, Y, F, H, L, I, and R yielded the lowest RMSE (1.49) between the predicted and crystal structure neighbor counts (Figures 1B and 2D). These residues are characterized by high to intermediate reactivity toward hydroxyl radicals (relative to proline) and are frequently present in experimental HRPF data sets. Our regression model generated more accurate neighbor count predictions without the need for molecular dynamics simulations or Rosetta mover models.

Implementation of an HRPF-Guided Scoring Term Improved Best-Scoring Model RMSD. We used the predicted neighbor counts from our regression model to implement an improved HRPF-based Rosetta score term. Our primary goal was to improve upon our previous implementations by generating more accurate expected neighbor counts for a larger subset of residue types. This would allow us to more accurately penalize decoys that do not agree with the experimental data. We opted to construct a sigmoidal scoring term that rewarded residues based on their agreement between their observed and predicted neighbor counts. Our previous scoring term used a deviation value of 3.5, as it was found to best differentiate between high RMSD and low RMSD models.[54] We hypothesized that our improved regression model would allow us to use a smaller value during rescoring. For our benchmark set ab initio models, we found that a deviation value of 2.1 was best able to distinguish between low RMSD and high RMSD models, indeed corroborating our hypothesis. This is likely because the

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

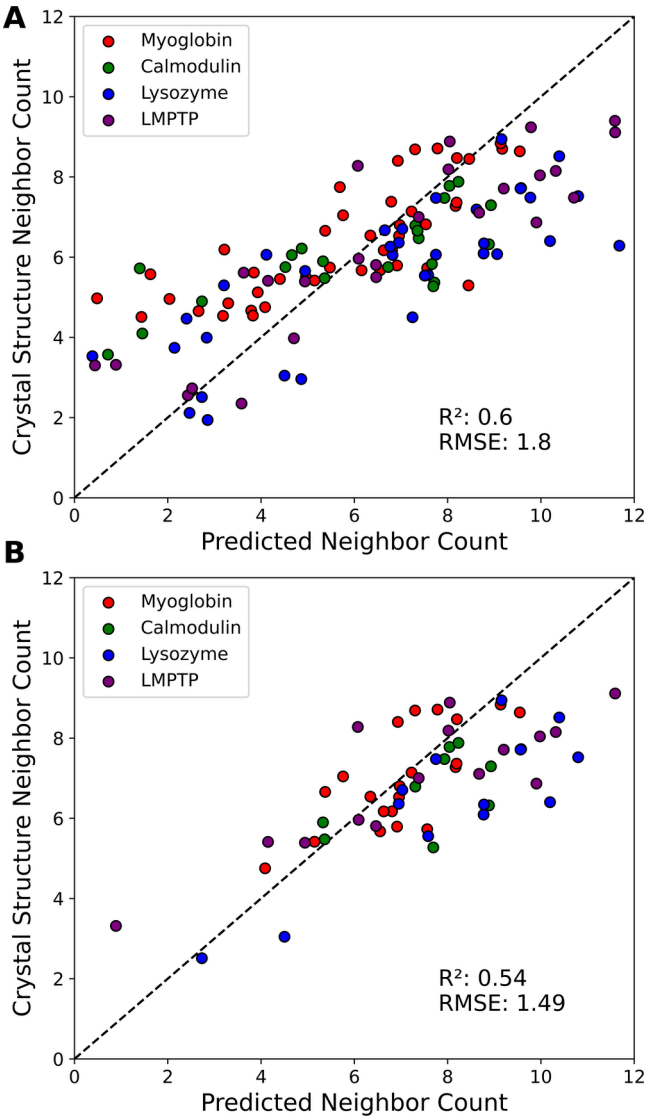


Figure 1. Correlation between the crystal structure and predicted neighbor counts for the benchmark set labeled residues. The $y = x$ line, R^2, and RMSE values are included in the plots. Residues are color coded by protein. (A) LightGBM model performance when predicting conical neighbor count for all labeled residues in the benchmark set. (B) LightGBM model performance when predicting the conical neighbor of labeled WYFHLIR residues in the benchmark set.

increased accuracy of our neighbor count predictions allowed us to impose stricter thresholds for agreement.

A smaller deviation value restricts the number of residues being scored, emphasizing residues that contribute the most useful structural information while minimizing the inclusion of labeling data for less relevant or misleading residues. We then applied our score term to a structure prediction task. We generated 20,000 ab initio models for each benchmark protein and rescored those by adding the *hrf_GBM* score term to the default Ref15 score as a weighted sum. Score versus RMSD to the crystal structure plots can be found in Figure 3: Rosetta Ref15 versus RMSD plots for all benchmark proteins can be found in Figure 3A, while (Ref15 + *hrf_GBM*) versus RMSD plots can be found in Figure 3B. Prior to rescoring, models with near-atomic-level detail (RMSD to crystal structure <5 Å) could only be identified for LMPTP. However, after rescoring, we were able to identify models with near-atomic-level detail for all four benchmark proteins. The best-scoring RMSD for myoglobin decreased from 6.61 to 1.95 Å, with the lowest RMSD model generated being 1.94 Å. Rescoring calmodulin with *hrf_GBM* resulted in the most dramatic improvement, with the lowest-scoring RMSD decreasing from 13.0 to 4.26 Å. We were able to select for the lowest RMSD model in our decoy set for lysozyme, with the best-scoring RMSD improving from 11.06 to 4.24 Å upon rescoring. The model selected for LMPTP remained at 1.67 Å. As this model already had atomic-level accuracy, it was reassuring to see that this did not change upon the inclusion of our score term. Restricting RMSD calculations to only residues that are involved in secondary structure elements (SSEs) resulted in similar improvements. The $RMSD_{SSE}$ for myoglobin improved from 6.20 to 1.41 Å after rescoring. Calmodulin best-scoring $RMSD_{SSE}$ improved from 12.7 to 3.8 Å, while the lysozyme $RMSD_{SSE}$ improved from 10.63 to 4.07 Å after rescoring. As the best-scoring LMPTP model did not change after rescoring, the $RMSD_{SSE}$ remained at 1.50 Å. The best-scoring models aligned to their respective crystal structures are shown in Figure 4: the best-scoring models prior to rescoring are shown in Figure 4A, while the best-scoring models after rescoring with *hrf_GBM* are shown in Figure 4B.

As the accuracy of our models was primarily assessed using alpha-carbon RMSD, we strove to find another metric that could quantitatively determine the effect that our score term had on side-chain orientation. We used the neighbor count deviation (ΔNC) between labeled residues in the best-scoring model and the crystal structure to quantify this effect. Because the conical neighbor count is sensitive to changes in both local residue packing and side-chain directionality, ΔNC provides a residue-specific measure of how closely the modeled side-chain environment reflects that of the native structure. Thus, improvements in ΔNC following rescoring indicate an improved side-chain placement for HRPF-labeled residues. Figure 5 shows the ΔNC distributions for the benchmark proteins prior to and after rescoring with *hrf_GBM*. The mean ΔNC for labeled myoglobin residues decreased from 1.10 to 0.62 after rescoring, while the mean ΔNC for labeled calmodulin residues decreased from 1.59 to 0.63. The mean ΔNC for labeled lysozyme residues decreased from 1.64 to 1.17, and the mean ΔNC for LMPTP residues remained at 0.76. These results further demonstrate that our approach improves accuracy on both a global and a residue level.

For further analysis, we decided to directly compare our benchmark set results to those published in our previous work.[54] As we used the same benchmark set of proteins in both cases, we were able to directly measure the impact of our improved neighbor count predictions and widened residue set on prediction accuracy. Figure 6 contains a comparison of the best-scoring model RMSD after rescoring (and, in the case of our previous method, after the generation of mover models). On average, our previous method yielded a best-scoring model RMSD of 3.88 Å after rescoring and application of mover models; our hybrid approach improved this to 3.03 Å. The largest improvement was seen for calmodulin, which improved from 7.24 to 4.26 Å. A similar improvement was observed when $RMSD_{SSE}$ was compared between methods. Our previous method yielded a mean $RMSD_{SSE}$ of 2.85 Å, while the current approach yielded a value of 2.63 Å. We also improved the labeled residue ΔNC for best-scoring models. The mean labeled residue ΔNC for the benchmark set proteins was 0.82 in our previous work. Our hybrid approach yielded a mean ΔNC of 0.74 when examining the same residue types. These results demonstrated

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

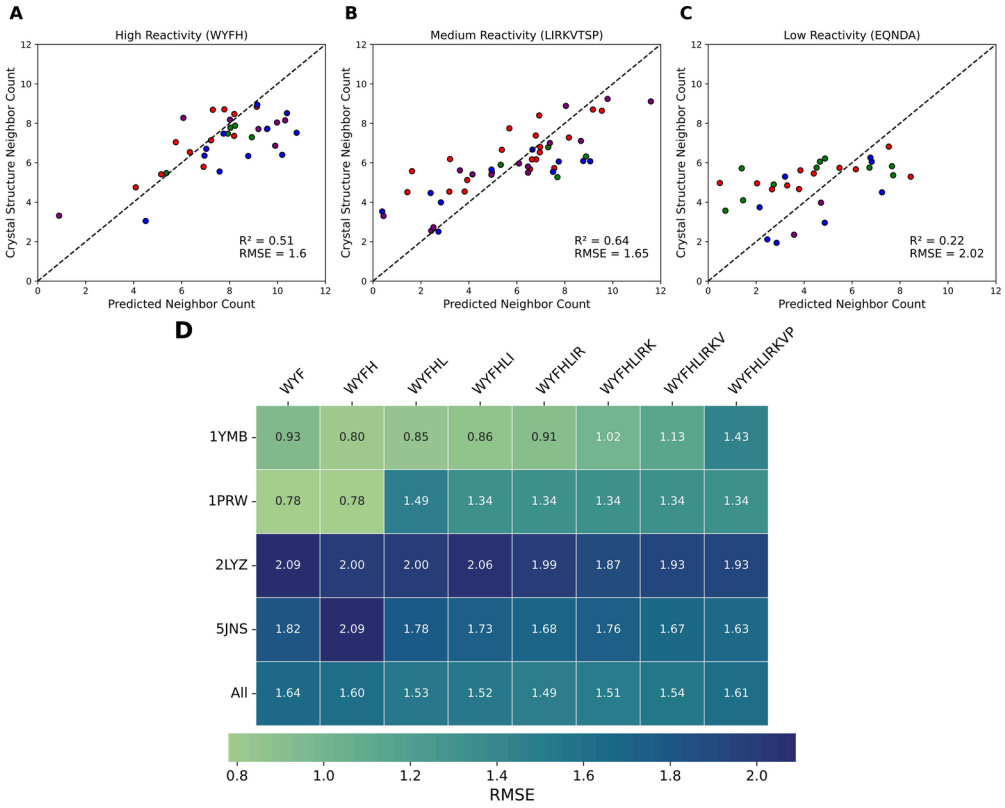


Figure 2. Myoglobin residues are in red, calmodulin residues in green, lysozyme residues in blue, and LMPTP residues in purple. (A) Crystal structure versus predicted neighbor count for highly reactive (>9.3× more reactive than Pro) residues. (B) Crystal structure versus predicted neighbor count for moderately reactive (1−9.3× more reactive than Pro) residues. (C) Crystal structure versus predicted neighbor count for residues with low (less reactive than Pro) reactivity. (D) RMSE heatmap for benchmark proteins. Each row corresponds to a different protein in the benchmark set, while each column corresponds to a different residue combination that was tested. The minimum RMSE in the "All" row corresponds to the WYFHLIR residue set, which was used for rescoring.

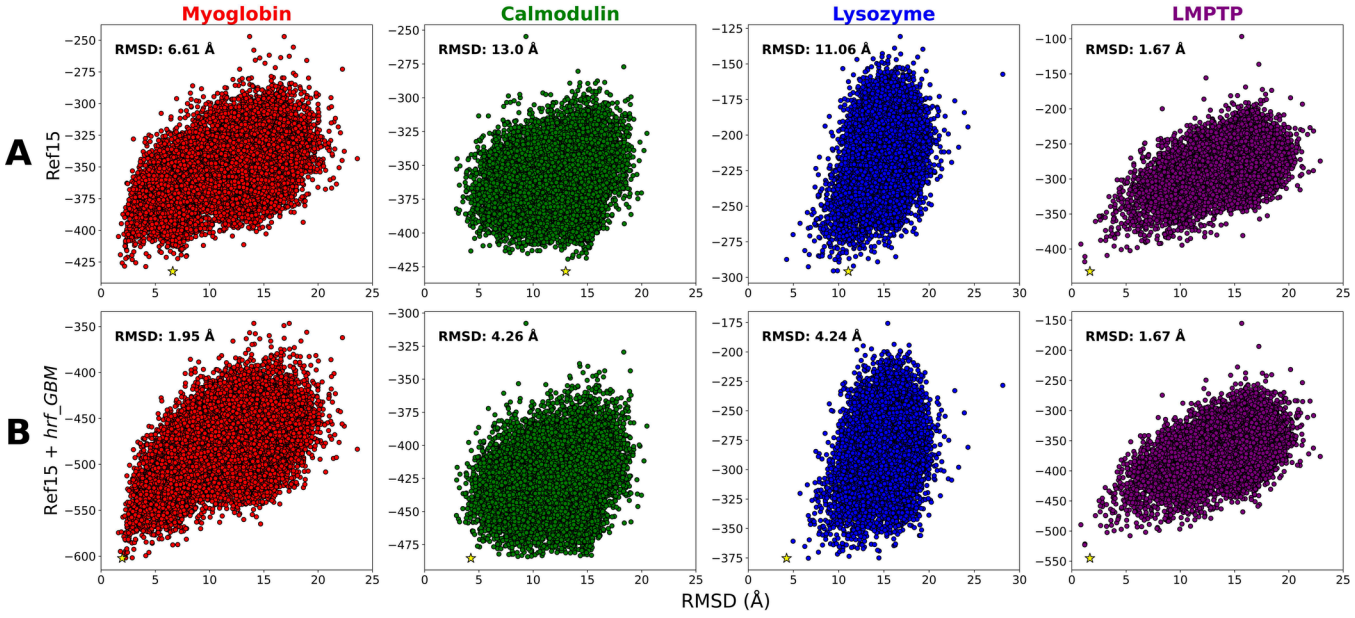


Figure 3. Score versus RMSD to the crystal structure plots for 20,000 ab initio models per benchmark protein. The lowest-scoring model for each benchmark protein is marked using a yellow star, and the corresponding model RMSD is included in the plot. (A) Rosetta Ref15 score versus RMSD. (B) Ref15 + *hrf_GBM* score term versus RMSD .

the utility of a hybrid structure prediction method and validated our hypothesis that machine learning models can be used to extract more high-quality structural information from experimental HRPF data.

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

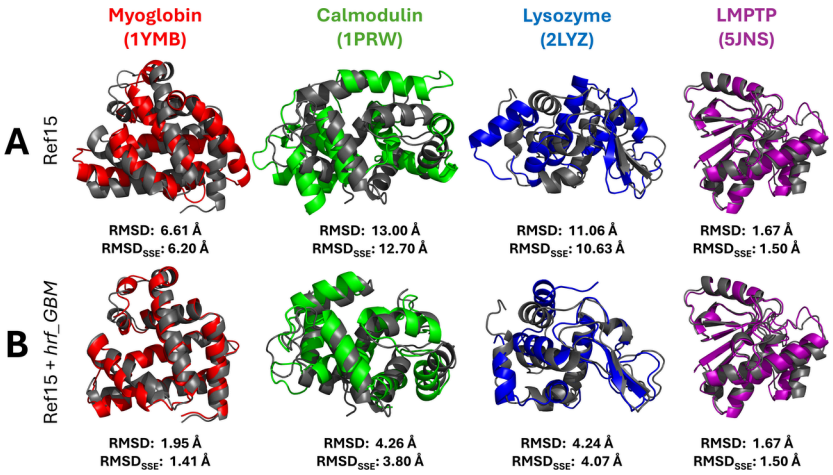


Figure 4. Best-scoring models (color) aligned to benchmark set crystal structures (gray). Models shown correspond to the predictions plotted with yellow stars in Figure 3. RMSD and RMSD$_{SSE}$ are included for each prediction. Myoglobin models are shown in red, calmodulin models in green, lysozyme models in blue, and LMPTP models in purple. (A) Best-scoring model based on Rosetta Ref15 score. (B) Best-scoring model after rescoring with *hrf_GBM*.

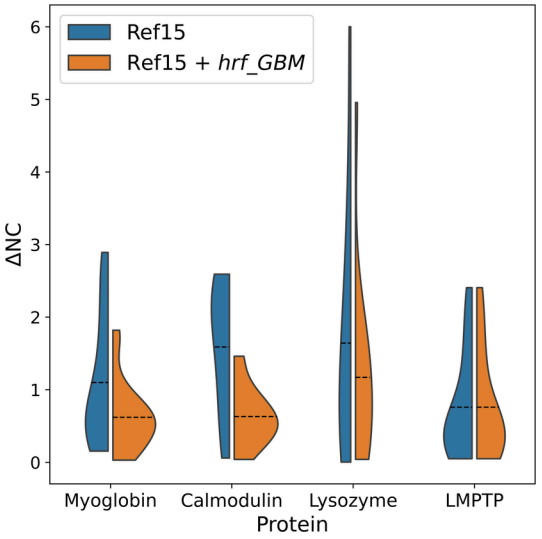


Figure 5. Violin plots showing the ΔNC values of labeled WYFHLIR residues in the best-scoring decoys for each benchmark protein. The mean ΔNCs are shown as dotted lines.

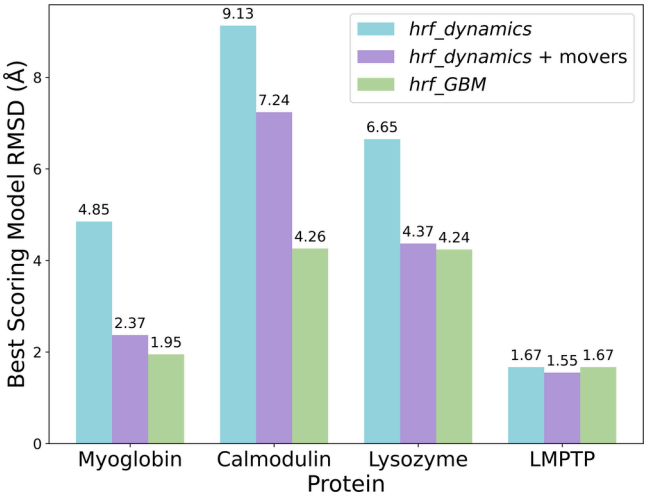


Figure 6. Grouped bar plots of the best-scoring model RMSD for each benchmark protein after rescoring with different HRPF-guided score terms. The green bars correspond to the aligned structures seen in Figure 4B.

Beyond enhancing the accuracy of top-scoring models, we also sought to evaluate whether our hybrid method could achieve these results with greater speed and lower computational cost in comparison to previous approaches. We benchmarked the speed of each method on the Owens cluster of the Ohio Supercomputer Center using Intel Xeon E5-2680 v4 CPUs at 2.4 GHz.[104] We opted to compare the speed of mover model generation to the speed of LightGBM feature generation, as these are the steps that take the longest. Mover ensembles were generated using 28 CPU cores and required about 5 h on average to generate the 600 mover models required for rescoring. The generation of LightGBM input features required approximately 0.5 h on the same system and number of CPU cores. Compared with our previous approach, our hybrid method resulted in a 10-fold decrease in runtime and computational cost. These improvements further emphasize the strength and utility of ML in structure prediction tasks.

CONCLUSION

In this study, we presented a hybrid structure prediction technique that combines elements of machine learning with Rosetta to further improve HRPF-guided prediction accuracy. Using experimental HRPF data, we trained a gradient boosting regression model to predict residue neighbor counts. We specifically chose this model architecture because gradient boosting models excel at capturing complex relationships between input features, allowing our model to extract richer structural insights from sparse HRPF data. We demonstrated that the inclusion of our novel HRPF-guided score term can yield accurate models for all four proteins in our benchmark set. Unlike our previous methods, our hybrid approach did not require the approximation of side-chain dynamics, which can be computationally expensive. The *hrf_GBM* score term can be used through the *hrf_GBM_rescore* application within Rosetta, and a pretrained model for predicting neighbor count can be found on GitHub: https://github.com/LindertLab/HRF_

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

GBM_Rescore. In-depth tutorials for generating the required input features and using the application can be found in the Supporting Information. In future work, we will focus on adapting our hybrid approach to work with other labeling reagents as well as protein complexes. We will also investigate whether machine learning can be used to predict HRPF labeling rates from structural features. Finally, our study also hints at the potential power of an AF2-like ML-based prediction network guided by experimental covalent labeling data. Despite the challenges associated with including experimental data in AF2, future work will explore this avenue.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.5c00053.

HRPF data used for model training and ab initio benchmark, tutorials, and additional figures (feature selection results from iterative cross-validation and feature importance analysis) (PDF)

Example input files and HRPF data used in tutorials (ZIP)

■ 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

Author

Elijah H. Day − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0009-0000-5927-4426

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.5c00053

Author Contributions

E.H.D. and S.L. conceived of the project. E.H.D. designed the neighbor count prediction model. E.H.D. and S.L. developed the Rosetta rescoring strategy, and E.H.D. performed the structure prediction benchmark. E.H.D. and S.L. wrote the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank the members of the Lindert group for their insightful discussions and suggestions. We would like to thank the Ohio Supercomputer Center[104] for valuable computational resources and technical support.

■ REFERENCES

(1) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93* (4), 1866−1879.

(2) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-Resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587* (7832), 157−161.

(3) Smyth, M. S.; Martin, J. H. J. X Ray Crystallography. *Molecular Pathology* **2000**, *53* (1), 8−14.

(4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(5) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493−500.

(6) Bertoline, L. M. F.; Lima, A. N.; Krieger, J. E.; Teixeira, S. K. Before and after AlphaFold2: An Overview of Protein Structure Prediction. *Front Bioinform* **2023**, *3*, No. 1120370.

(7) Konermann, L.; Vahidi, S.; Sowole, M. A. Mass Spectrometry Methods for Studying Structure and Dynamics of Biological Macromolecules. *Anal. Chem.* **2014**, *86* (1), 213−232.

(8) Britt, H. M.; Cragnolini, T.; Thalassinos, K. Integration of Mass Spectrometry Data for Structural Biology. *Chem. Rev.* **2022**, *122* (8), 7952−7986.

(9) Banazadeh, A.; Veillon, L.; Wooding, K. M.; Zabet-moghaddam, M.; Mechref, Y. Recent Advances in Mass Spectrometric Analysis of Glycoproteins. *ELECTROPHORESIS* **2017**, *38* (1), 162−189.

(10) Barrera, N. P.; Zhou, M.; Robinson, C. V. The Role of Lipids in Defining Membrane Protein Interactions: Insights from Mass Spectrometry. *Trends in Cell Biology* **2013**, *23* (1), 1−8.

(11) O'Reilly, F. J.; Rappsilber, J. Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology. *Nat. Struct Mol. Biol.* **2018**, *25* (11), 1000−1008.

(12) 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* (7), 595−602.

(13) Konermann, L.; Pan, J.; Liu, Y.-H. Hydrogen Exchange Mass Spectrometry for Studying Protein Structure and Dynamics. *Chem. Soc. Rev.* **2011**, *40* (3), 1224−1234.

(14) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. The Power of Ion Mobility-Mass Spectrometry for Structural Characterization and the Study of Conformational Dynamics. *Nat. Chem.* **2014**, *6* (4), 281−294.

(15) Snyder, D. T.; Harvey, S. R.; Wysocki, V. H. Surface-Induced Dissociation Mass Spectrometry as a Structural Biology Tool. *Chem. Rev.* **2022**, *122* (8), 7442−7487.

(16) Limpikirati, P.; Liu, T.; Vachet, R. W. Covalent Labeling-Mass Spectrometry with Non-Specific Reagents for Studying Protein Structure and Interactions. *Methods* **2018**, *144*, 79−93.

(17) Bamberger, C.; Pankow, S.; Martínez-Bartolomé, S.; Ma, M.; Diedrich, J.; Rissman, R. A.; Yates, J. R. Protein Footprinting via Covalent Protein Painting Reveals Structural Changes of the Proteome in Alzheimer's Disease. *J. Proteome Res.* **2021**, *20* (5), 2762−2771.

(18) Liu, X. R.; Zhang, M. M.; Gross, M. L. Mass Spectrometry-Based Protein Footprinting for Higher-Order Structure Analysis: Fundamentals and Applications. *Chem. Rev.* **2020**, *120* (10), 4355−4454.

(19) Manzi, L.; Barrow, A. S.; Scott, D.; Layfield, R.; Wright, T. G.; Moses, J. E.; Oldham, N. J. Carbene Footprinting Accurately Maps

Binding Sites in Protein−Ligand and Protein−Protein Interactions. *Nat. Commun.* **2016**, *7* (1), No. 13288.
(20) Limpikirati, P.; Pan, X.; Vachet, R. W. Covalent Labeling with Diethylpyrocarbonate: Sensitive to the Residue Microenvironment, Providing Improved Analysis of Protein Higher Order Structure by Mass Spectrometry. *Anal. Chem.* **2019**, *91* (13), 8516−8523.
(21) Tremblay, C. Y.; Kirsch, Z. J.; Vachet, R. W. Epitope Mapping with Diethylpyrocarbonate Covalent Labeling-Mass Spectrometry. *Anal. Chem.* **2022**, *94* (2), 1052−1059.
(22) Imiołek, M.; Karunanithy, G.; Ng, W.-L.; Baldwin, A. J.; Gouverneur, V.; Davis, B. G. Selective Radical Trifluoromethylation of Native Residues in Proteins. *J. Am. Chem. Soc.* **2018**, *140* (5), 1568−1571.
(23) Xu, G.; Chance, M. R. Hydroxyl Radical-Mediated Modification of Proteins as Probes for Structural Proteomics. *Chem. Rev.* **2007**, *107* (8), 3514−3543.
(24) Kiselar, J. G.; Maleknia, S. D.; Sullivan, M.; Downard, K. M.; Chance, M. R. Hydroxyl Radical Probe of Protein Surfaces Using Synchrotron X-Ray Radiolysis and Mass Spectrometry. *Int. J. Radiat. Biol.* **2002**, *78* (2), 101−114.
(25) Johnson, D. T.; Di Stefano, L. H.; Jones, L. M. Fast Photochemical Oxidation of Proteins (FPOP): A Powerful Mass Spectrometry−Based Structural Proteomics Tool. *J. Biol. Chem.* **2019**, *294* (32), 11969−11979.
(26) McKenzie-Coe, A.; Montes, N. S.; Jones, L. M. Hydroxyl Radical Protein Footprinting: A Mass Spectrometry-Based Structural Method for Studying the Higher Order Structure of Proteins. *Chem. Rev.* **2022**, *122* (8), 7532−7561.
(27) Li, Z.; Moniz, H.; Wang, S.; Ramiah, A.; Zhang, F.; Moremen, K. W.; Linhardt, R. J.; Sharp, J. S. High Structural Resolution Hydroxyl Radical Protein Footprinting Reveals an Extended Robo1-Heparin Binding Interface. *J. Biol. Chem.* **2015**, *290* (17), 10729−10740.
(28) Li, X.; Grant, O. C.; Ito, K.; Wallace, A.; Wang, S.; Zhao, P.; Wells, L.; Lu, S.; Woods, R. J.; Sharp, J. S. Structural Analysis of the Glycosylated Intact HIV-1 Gp120-B12 Antibody Complex Using Hydroxyl Radical Protein Footprinting. *Biochemistry* **2017**, *56* (7), 957−970.
(29) Watson, C.; Sharp, J. S. Conformational Analysis of Therapeutic Proteins by Hydroxyl Radical Protein Footprinting. *AAPS J.* **2012**, *14* (2), 206−217.
(30) Ralston, C. Y.; Sharp, J. S. Structural Investigation of Therapeutic Antibodies Using Hydroxyl Radical Protein Footprinting Methods. *Antibodies (Basel)* **2022**, *11* (4), 71.
(31) Aprahamian, M. L.; Lindert, S. Utility of Covalent Labeling Mass Spectrometry Data in Protein Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2019**, *15* (5), 3410−3424.
(32) Biehn, S. E.; Lindert, S. Protein Structure Prediction with Mass Spectrometry Data. *Annu. Rev. Phys. Chem.* **2022**, *73* (1), 1−19.
(33) Biehn, S. E.; Picarello, D. M.; Pan, X.; Vachet, R. W.; Lindert, S. Accounting for Neighboring Residue Hydrophobicity in Diethylpyrocarbonate Labeling Mass Spectrometry Improves Rosetta Protein Structure Prediction. *J. Am. Soc. Mass Spectrom.* **2022**, *33* (3), 584−591.
(34) Seffernick, J. T.; Canfield, S. M.; Harvey, S. R.; Wysocki, V. H.; Lindert, S. Prediction of Protein Complex Structure Using Surface-Induced Dissociation and Cryo-Electron Microscopy. *Anal. Chem.* **2021**, *93* (21), 7596−7605.
(35) Turzo, S. M. B. A.; Seffernick, J. T.; Rolland, A. D.; Donor, M. T.; Heinze, S.; Prell, J. S.; Wysocki, V. H.; Lindert, S. Protein Shape Sampled by Ion Mobility Mass Spectrometry Consistently Improves Protein Structure Prediction. *Nat. Commun.* **2022**, *13* (1), 4377.
(36) Drake, Z. C.; Seffernick, J. T.; Lindert, S. Protein Complex Prediction Using Rosetta, AlphaFold, and Mass Spectrometry Covalent Labeling. *Nat. Commun.* **2022**, *13* (1), 7846.
(37) Kahraman, A.; Herzog, F.; Leitner, A.; Rosenberger, G.; Aebersold, R.; Malmström, L. Cross-Link Guided Molecular Modeling with ROSETTA. *PLoS One* **2013**, *8* (9), No. e73411.
(38) 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* (24), 15709−15717.
(39) Biehn, S. E.; Limpikirati, P.; Vachet, R. W.; Lindert, S. Utilization of Hydrophobic Microenvironment Sensitivity in Diethylpyrocarbonate Labeling for Protein Structure Prediction. *Anal. Chem.* **2021**, *93* (23), 8188−8195.
(40) Seffernick, J. T.; Lindert, S. Hybrid Methods for Combined Experimental and Computational Determination of Protein Structure. *J. Chem. Phys.* **2020**, *153* (24), No. 240901.
(41) Turzo, S. M. B. A.; Seffernick, J. T.; Lyskov, S.; Lindert, S. Predicting Ion Mobility Collision Cross Sections Using Projection Approximation with ROSIE-PARCS Webserver. *Brief Bioinform* **2023**, *24* (5), No. bbad308.
(42) Nguyen, T. T.; Marzolf, D. R.; Seffernick, J. T.; Heinze, S.; Lindert, S. Protein Structure Prediction Using Residue-Resolved Protection Factors from Hydrogen-Deuterium Exchange NMR. *Structure* **2022**, *30* (2), 313.
(43) Marzolf, D. R.; Seffernick, J. T.; Lindert, S. Protein Structure Prediction from NMR Hydrogen-Deuterium Exchange Data. *J. Chem. Theory Comput* **2021**, *17* (4), 2619−2629.
(44) 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* (8), 1330−1341.
(45) 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* (17), 8143−8148.
(46) 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* (12), 7721−7729.
(47) 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. *Sci. Rep* **2017**, *7* (1), 4552.
(48) Khaje, N. A.; Eletsky, A.; Biehn, S. E.; Mobley, C. K.; Rogals, M. J.; Kim, Y.; Mishra, S. K.; Doerksen, R. J.; Lindert, S.; Prestegard, J. H.; Sharp, J. S. Validated Determination of NRG1 Ig-like Domain Structure by Mass Spectrometry Coupled with Computational Modeling. *Commun. Biol.* **2022**, *5* (1), 452.
(49) Seffernick, J. T.; Turzo, S. B. A.; Harvey, S. R.; Kim, Y.; Somogyi, Á.; Marciano, S.; Wysocki, V. H.; Lindert, S. Simulation of Energy-Resolved Mass Spectrometry Distributions from Surface-Induced Dissociation. *Anal. Chem.* **2022**, *94* (29), 10506−10514.
(50) Lössl, P.; Kölbel, K.; Tänzler, D.; Nannemann, D.; Ihling, C. H.; Keller, M. V.; Schneider, M.; Zaucke, F.; Meiler, J.; Sinz, A. Analysis of Nidogen-1/Laminin Γ1 Interaction by Cross-Linking, Mass Spectrometry, and Computational Modeling Reveals Multiple Binding Modes. *PLoS One* **2014**, *9* (11), No. e112886.
(51) 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. *J. Struct. Biol.* **2016**, *196* (3), 340−349.
(52) Khakzad, H.; Happonen, L.; Malmström, J.; Malmström, L. Cheetah-MS: A Web Server to Model Protein Complexes Using Tandem Cross-Linking Mass Spectrometry Data. *Bioinformatics* **2021**, *37* (24), 4871−4872.
(53) 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* (1), 192.
(54) Biehn, S. E.; Lindert, S. Accurate Protein Structure Prediction with Hydroxyl Radical Protein Footprinting Data. *Nat. Commun.* **2021**, *12* (1), 341.
(55) Ahdritz, G.; Bouatta, N.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; Nowaczynski, A.; Wang, B.; Stepniewska-Dziubinska, M. M.; Zhang, S.;

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

Ojewole, A.; Guney, M. E.; Biderman, S.; Watkins, A. M.; Ra, S.; Lorenzo, P. R.; Nivon, L.; Weitzner, B.; Ban, Y.-E. A.; Sorger, P. K.; Mostaque, E.; Zhang, Z.; Bonneau, R.; AlQuraishi, M. OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization. *bioRxiv* **2022**, 1.

(56) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. Protein Structure Prediction with In-Cell Photo-Crosslinking Mass Spectrometry and Deep Learning. *Nat. Biotechnol.* **2023**, *41*, 1810.

(57) Stahl, K.; Warneke, R.; Demann, L.; Bremenkamp, R.; Hormes, B.; Brock, O.; Stülke, J.; Rappsilber, J. Modelling Protein Complexes with Crosslinking Mass Spectrometry and Deep Learning. *Nat. Commun.* **2024**, *15* (1), 7866.

(58) Drake, Z. C.; Day, E.; Toth, P.; Lindert, S. Deep-Learning Structure Elucidation from Single-Mutant Deep Mutational Scanning. *bioRxiv* **2024**, 1.

(59) Huang, W.; Ravikumar, K. M.; Chance, M. R.; Yang, S. Quantitative Mapping of Protein Structure by Hydroxyl Radical Footprinting-Mediated Structural Mass Spectrometry: A Protection Factor Analysis. *Biophys. J.* **2015**, *108* (1), 107−115.

(60) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. *30*.

(61) Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A Comparative Analysis of Gradient Boosting Algorithms. *Artif Intell Rev.* **2021**, *54* (3), 1937−1967.

(62) Gustavsson, M.; Wang, L.; van Gils, N.; 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. *Nat. Commun.* **2017**, *8* (1), No. 14135.

(63) Gau, B. C.; Chen, J.; Gross, M. L. Fast Photochemical Oxidation of Proteins for Comparing Solvent-Accessibility Changes Accompanying Protein Folding: Data Processing and Application to Barstar. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2013**, *1834* (6), 1230−1238.

(64) Gupta, S.; Mangel, W. F.; McGrath, W. J.; Perek, J. L.; Lee, D. W.; Takamoto, K.; Chance, M. R. DNA Binding Provides a Molecular Strap Activating the Adenovirus Proteinase. *Molecular & Cellular Proteomics* **2004**, *3* (10), 950−959.

(65) Zheng, X.; Wintrode, P. L.; Chance, M. R. Complementary Structural Mass Spectrometry Techniques Reveal Local Dynamics in Functionally Important Regions of a Metastable Serpin. *Structure* **2008**, *16* (1), 38−51.

(66) Xu, M.; Kiselar, J.; Whited, T. L.; Hernandez-Sanchez, W.; Taylor, D. J. POT1-TPP1 Differentially Regulates Telomerase via POT1 His266 and as a Function of Single-Stranded Telomere DNA Length. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (47), 23527−23533.

(67) Chea, E. E.; Deredge, D. J.; Jones, L. M. Insights on the Conformational Ensemble of Cyt C Reveal a Compact State during Peroxidase Activity. *Biophys. J.* **2020**, *118* (1), 128−137.

(68) Ruan, H.; Kiselar, J.; Zhang, W.; Li, S.; Xiong, R.; Liu, Y.; Yang, S.; Lai, L. Integrative Structural Modeling of a Multidomain Polo-like Kinase. *Phys. Chem. Chem. Phys.* **2020**, *22* (47), 27581−27589.

(69) Cornwell, O.; Radford, S. E.; Ashcroft, A. E.; Ault, J. R. Comparing Hydrogen Deuterium Exchange and Fast Photochemical Oxidation of Proteins: A Structural Characterisation of Wild-Type and ΔN6 B2-Microglobulin. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (12), 2413−2426.

(70) Hamid, M.; Khalid, M. F.; Chaudhary, S. U.; Khan, S. The Solvation of the E. Coli CheY Phosphorylation Site Mapped by XFMS. *International Journal of Molecular Sciences* **2022**, *23* (21), 12771.

(71) Kiselar, J. G.; Janmey, P. A.; Almo, S. C.; Chance, M. R. Structural Analysis of Gelsolin Using Synchrotron Protein Footprinting. *Mol. Cell Proteomics* **2003**, *2* (10), 1120−1132.

(72) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403−410.

(73) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(74) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (22), 10915−10919.

(75) Henikoff, S.; Henikoff, J. G. Position-Based Sequence Weights. *J. Mol. Biol.* **1994**, *243* (4), 574−578.

(76) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED Protein Structure Prediction Server. *Bioinformatics* **2000**, *16* (4), 404−405.

(77) Ward, J. J.; McGuffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T. The DISOPRED Server for the Prediction of Protein Disorder. *Bioinformatics* **2004**, *20* (13), 2138−2139.

(78) Bowie, J. U.; Lüthy, R.; Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **1991**, *253* (5016), 164−170.

(79) Fan, C.; Liu, D.; Huang, R.; Chen, Z.; Deng, L. PredRSA: A Gradient Boosted Regression Trees Approach for Predicting Protein Solvent Accessibility. *BMC Bioinformatics* **2016**, *17* (1), S8.

(80) Li, N.; Sun, Z.; Jiang, F. Prediction of Protein-Protein Binding Site by Using Core Interface Residue and Support Vector Machine. *BMC Bioinformatics* **2008**, *9* (1), 553.

(81) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235.

(82) Hubbard, S. J.; Thornton, J. M. *NACCESS*; Computer Program, 1993.

(83) Ziemianowicz, D. S.; MacCallum, J. L.; Schriemer, D. C. Correlation between Labeling Yield and Surface Accessibility in Covalent Labeling Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (2), 207−216.

(84) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J.* **2015**, *7* (1), 4−14.

(85) Cruciani, G.; Baroni, M.; Carosati, E.; Clementi, M.; Valigi, R.; Clementi, S. Peptide Studies by Means of Principal Properties of Amino Acids Derived from MIF Descriptors. *Journal of Chemometrics* **2004**, *18* (3−4), 146−155.

(86) Liang, G.; Chen, G.; Niu, W.; Li, Z. Factor Analysis Scales of Generalized Amino Acid Information as Applied in Predicting Interactions between the Human Amphiphysin-1 SH3 Domains and Their Peptide Ligands. *Chem. Biol. Drug Des* **2008**, *71* (4), 345−351.

(87) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. *J. Protein Chem.* **1985**, *4* (1), 23−55.

(88) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481−2491.

(89) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs. *Biopolymers* **2005**, *80* (6), 775−786.

(90) Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830* (1), 106−115.

(91) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38* (3), 805−816.

(92) van Westen, G. J.; Swier, R. F.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *Journal of Cheminformatics* **2013**, *5* (1), 41.

(93) Georgiev, A. G. Interpretable Numerical Descriptors of Amino Acid Space. *J. Comput. Biol.* **2009**, *16* (5), 703−723.

(94) Magnusson, M.; Vehtari, A.; Jonasson, J.; Andersen, M. Leave-One-Out Cross-Validation for Bayesian Model Comparison in Large Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*; PMLR, 2020; pp 341−351.

(95) Kaur, P.; Kiselar, J.; Yang, S.; Chance, M. R. Quantitative Protein Topography Analysis and High-Resolution Structure Prediction Using

https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347

Hydroxyl Radical Labeling and Tandem-Ion Mass Spectrometry (MS)*. *Molecular & Cellular Proteomics* **2015**, *14* (4), 1159−1168.
(96) Stanford, S. M.; Aleshin, A. E.; Zhang, V.; Ardecky, R. J.; Hedrick, M. P.; Zou, J.; Ganji, S. R.; Bliss, M. R.; Yamamoto, F.; Bobkov, A. A.; Kiselar, J.; Liu, Y.; Cadwell, G. W.; Khare, S.; Yu, J.; Barquilla, A.; Chung, T. D.; Mustelin, T.; Schenk, S.; Bankston, L. A.; Liddington, R. C.; Pinkerton, A. B.; Bottini, N. Diabetes Reversal by Inhibition of the Low Molecular Weight Tyrosine Phosphatase. *Nat. Chem. Biol.* **2017**, *13* (6), 624−632.
(97) Bender, B. J.; Cisneros, A. I.; 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.
(98) 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. *J. Chem. Theory Comput* **2017**, *13* (6), 3031−3048.
(99) Kim, D. E.; Chivian, D.; Baker, D. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Res.* **2004**, *32*, W526−W531.
(100) Maleknia, S. D.; Downard, K. M. Advances in Radical Probe Mass Spectrometry for Protein Footprinting in Chemical Biology Applications. *Chem. Soc. Rev.* **2014**, *43* (10), 3244−3258.
(101) Zhao, Y.; Zhang, W.; Hong, J.; Yang, L.; Wang, Y.; Qu, F.; Xu, W. Mobility Capillary Electrophoresis−Native Mass Spectrometry Reveals the Dynamic Conformational Equilibrium of Calmodulin and Its Complexes. *Analyst* **2024**, *149* (14), 3793−3802.
(102) Vigil, D.; Gallagher, S. C.; Trewhella, J.; García, A. E. Functional Dynamics of the Hydrophobic Cleft in the N-Domain of Calmodulin. *Biophys. J.* **2001**, *80* (5), 2082−2092.
(103) Grabarek, Z. Structural Basis for Diversity of the EF-Hand Calcium-Binding Proteins. *J. Mol. Biol.* **2006**, *359* (3), 509−525.
(104) Ohio Supercomputer Center. 1987. osc.edu/ark:/19495/f5s1ph73.



https://doi.org/10.1021/jasms.5c00053
J. Am. Soc. Mass Spectrom. 2025, 36, 1336−1347