analytical chemistry

pubs.acs.org/ac

Article

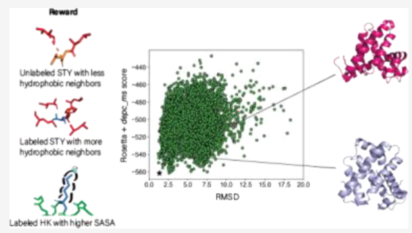
Utilization of Hydrophobic Microenvironment Sensitivity in Diethylpyrocarbonate Labeling for Protein Structure Prediction

Sarah E. Biehn, Patanachai Limpikirati, Richard W. Vachet, and Steffen Lindert*

Cite This: *Anal. Chem.* 2021, 93, 8188–8195

Read Online

ABSTRACT: Diethylpyrocarbonate (DEPC) labeling analyzed with mass spectrometry can provide important insights into higher order protein structures. It has been previously shown that neighboring hydrophobic residues promote a local increase in DEPC concentration such that serine, threonine, and tyrosine residues are more likely to be labeled despite low solvent exposure. In this work, we developed a Rosetta algorithm that used the knowledge of labeled and unlabeled serine, threonine, and tyrosine residues and assessed their local hydrophobic environment to improve protein structure prediction. Additionally, DEPC-labeled histidine and lysine residues with higher relative solvent accessible surface area values (i.e., more exposed) were scored favorably. Application of our score term led to reductions of the root-mean-square deviations (RMSDs) of the lowest scoring models. Additionally, models that scored well tended to have lower RMSDs. A detailed tutorial describing our protocol and required command lines is included. Our work demonstrated the considerable potential of DEPC covalent labeling data to be used for accurate higher order structure determination.



With the aid of various labeling reagents, mass spectrometry (MS) is emerging as an attractive technique for investigating protein structures. Techniques such as hydrogen-deuterium exchange, chemical cross-linking, and covalent labeling have been successfully employed to elucidate protein structures and dynamics.[1−3] Covalent labeling with MS (CL-MS) in which labeling reagents irreversibly modify protein residues can provide an insight into relative solvent exposure of labeled residues. Hydroxyl radical footprinting, radical trifluoromethylation, and carbene footprinting are promising techniques in covalent labeling mass spectrometry that rely upon label generation via photolysis or radiolysis.[4−6] Diethylpyrocarbonate (DEPC) is a popular covalent labeling reagent that is commercially available; it also does not require additional steps such as radical generation.[7−9] One advantage of DEPC as a labeling reagent is the single product generation for labeled residues. DEPC reacts with six nucleophilic residues (Cys, Lys, His, Ser, Thr, and Tyr) in addition to the protein N-terminus.[7,10,11] Labeled residues are identified by a mass increase of +72.021 Da.[12] Structures of DEPC and DEPC-modified residues are shown in Supplementary Figure 1. With careful attention to concentration and exposure times to avoid labeling-induced structural perturbation, cysteine scrambling,[13] or hydrolysis leading to label loss,[14] DEPC is a promising covalent labeling reagent to use for structure elucidation. While DEPC labeling yields valuable structural information, labeling data are too sparse to unambiguously determine the protein structure. Computational methods are necessary in combination with the DEPC labeling data in order to illuminate additional structural detail.

MS-guided modeling has previously been executed for protein structure investigation.[3,15−23] One tool that has been applied is Rosetta, a powerful molecular modeling software suite.[24,25] Among its many applications, structure prediction using sparse data from a variety of experimental techniques (including mass spectrometry) has been implemented.[24,26−34] Rosetta software has structural modeling applications, such as ab initio modeling that relies only on an amino acid sequence for structure building and template-guided homology modeling that can predict the protein structure.[25,35] Additionally, Rosetta is capable of assessing relative solvent exposure, making it an ideal tool to predict the protein structure from covalent labeling data. Functionality to use covalent labeling data to guide protein tertiary structure prediction has successfully been incorporated into Rosetta software and has shown to improve the model and distribution quality.[26,27,29] Overall, there is a growing need for automated and reliable algorithms that generate protein structures based on CL-MS data so as to move beyond reliance on manual interpretation of data in light of crystal structures and homology models.

Despite its ability to dissolve in aqueous solutions up to 40 mM, DEPC is a hydrophobic molecule with limited water

Received: January 27, 2021
Accepted: May 19, 2021
Published: June 1, 2021

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188–8195

solubility.[12,36] Recently, it was shown that protein microenvironments enriched in neighboring hydrophobic residues led to enhanced labeling efficiency of Ser, Thr, and Tyr (STY) residues. It was proposed that nearby hydrophobic residues were facilitating an increased local concentration of DEPC, thus making STY residues more likely to be labeled.[12] Here, we exploited the connection between the microenvironmental effect of neighboring hydrophobic residues and labeling of STY residues from DEPC-based CL-MS for structural modeling. We developed a score term to assess models based on relative solvent accessible surface area (SASA) values and hydrophobic neighbor counts for labeled and unlabeled STY residues. Additionally, as more exposed residues are more likely to be covalently labeled, our score term rewarded models with labeled histidine and lysine residues that exhibited higher relative SASA values. While covalent labeling data can be difficult to accurately quantify,[37,38] we have implemented a score term that relies only upon the residue DEPC-label status for structure prediction improvement. This is the first implementation of DEPC labeling data-guided structure prediction into the Rosetta software suite. When testing our algorithm on a benchmark set of six proteins, we found that inclusion of DEPC data led to lower, improved top scoring model root-mean-square deviation (RMSD) values and an improved funnel-like quality of the model distributions.

■ MATERIALS AND METHODS

Benchmark Set. The benchmark set comprised six proteins for which we obtained DEPC labeling data for His, Lys, Ser, Thr, and Tyr residues. The benchmark set included carbonic anhydrase (PDB 1V9E, 259 residues), ubiquitin (PDB 1UBQ, 76 residues), myoglobin (PDB 1DWR, 152 residues), β2-microglobulin (PDB 1JNJ, 100 residues), lysozyme (PDB 2LYZ, 129 residues), and human growth hormone (PDB 1HGU, 191 residues).

The DEPC labeling experiments and associated liquid chromatography (LC)-MS measurements were conducted as described in previous work.[12,14,39] For all the DEPC−protein reactions, conditions were chosen to achieve modification levels between 1 and 1.5 labels per protein on average to maintain the structural integrity of the protein.[7,40] Each protein was dissolved at a defined concentration between 10 and 50 μM in a 10 mM 3-(*N*-morpholino)propanesulfonic acid (MOPS) buffer at pH 7.4. Then, a 4- or 5-molar excess of DEPC was added and allowed to react for either 1 or 5 min. The reactions were performed at 37 °C and were quenched by the addition of imidazole. Between three and five replicates of the labeling experiments were carried out for each protein.

To enable identification of the DEPC-modified residues, the labeled proteins were digested using immobilized trypsin or chymotrypsin after buffer exchange into a phosphate buffer at pH 8.0. For proteins with disulfide bonds, reduction and alkylation with a 40-fold excess of tris(2-carboxyethyl)-phosphine (TCEP) and an 80-fold excess of iodoacetamide, respectively, were performed prior to digestion. Further experimental details about the digestions of ubiquitin,[12] β2-microglobulin,[12] human growth hormone,[12] carbonic anhydrase,[39] myoglobin,[39] and lysozyme[41] can be found elsewhere.

The peptide fragments after protein digestion were measured by LC/MS on a Thermo Scientific (Waltham, MA) Orbitrap Fusion mass spectrometer equipped with a nanoelectrospray ionization source. Online LC separations were conducted using a Thermo Scientific Easy-NanoLC 1000 system with a Thermo Scientific Acclaim PepMap C18 nanocolumn (15 cm × 75 μm ID, 2 μm, 100 Å). Peptides were eluted using a gradient of acetonitrile containing 0.1% formic acid at a flow rate of 0.3 μL/min. The gradient of acetonitrile was increased from 0 to 50% for 50 or 60 min, depending on the protein digest, before ramping up the acetonitrile to 100% for 15 additional minutes. The longer acetonitrile gradient was used for the digests of carbonic anhydrase, lysozyme, and myoglobin, while the shorter gradient was used for ubiquitin, β2-microglobulin, and human growth hormone. Peptides were identified and their DEPC labeling extents were determined using a custom software pipeline[42] that allows labeling percentages as low as 0.001% to be determined. Residue level DEPC modification percentages (% labeling) were obtained from chromatographic peak areas of the unmodified and modified peptides using approaches described previously.[39] In the work described here, a residue was considered labeled if its labeling percentages exceeded 0.01%.[43]

Ab Initio and Homology Model Generation with Rosetta. Fragment libraries were generated using the Robetta server for all six benchmark proteins.[44] 3mer and 9mer fragments and FASTA sequences of each benchmark proteins were used with the Rosetta *AbInitioRelax* protocol to generate 10,000 models per benchmark protein. The RMSD was calculated by supplying the respective crystal structures during scoring with the Rosetta energy function (abbreviated Ref15). Models were then ranked by score. The lowest RMSD model generated was used to determine if homology modeling was necessary for the particular protein. Proteins whose lowest RMSD model exhibited an RMSD greater than 5 Å were further modeled with homology modeling.

Homology models for carbonic anhydrase, lysozyme, and human growth hormone were generated with the Rosetta Comparative Modeling protocol.[35] For each protein, seven templates (Supplementary Table 1) with varying sequence coverage (60−100%) and identity (24−99%) were used during modeling. Each template was used for the generation of 500 models for a total of 3500 models built for each protein. Upon generation, models were relaxed with Rosetta's Relax application prior to scoring with the Ref15 and RMSD calculations.[45]

Identification of Hydrophobic Neighbor Count and Relative SASA Parameters. In order to derive the values used in the score term, we developed a custom Python script to identify the hydrophobic neighboring residues of labeled and unlabeled STY residues using the benchmark crystal structures. Crystal structures were only used for the initial derivation of score term parameters. Labeling data for STY residues is included in Supplementary Table 2. Distance (dist_{ij}) was calculated between the hydroxyl group oxygen atom in labeled and unlabeled STY residues (i) and the beta carbon in the side chain of hydrophobic residues (j). Hydrophobic neighboring residues were considered to be residue types Phe, Ile, Trp, Leu, Val, Met, Tyr, Ala, and Pro, as used previously.[12] The total contribution to the neighbor count was calculated as shown in eq 1.

$$\mathrm{hnc}_i = \sum_{i \neq j}^{\#\text{hydrophobic residues}} \frac{1.0}{1 + \exp(2 \times (\mathrm{dist}_{ij} - 8\ \text{Å})} \tag{1}$$

A midpoint value of 8.0 Å and a steepness value of 2.0 were chosen to give a full neighbor contribution up to distances of 6

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195

Å, the molecular dimensions of the DEPC molecule.[12] The relative SASA, the SASA of the residue sidechain normalized by the free residue SASA of the side chain, was calculated for the crystal structures using Rosetta RelSASA. Relative SASA values ranged from 0% indicating complete burial to 100% implying full exposure. A relative SASA range of 5−35% demonstrated ~1 residue difference in the average neighbor count between 24 labeled and 22 unlabeled residues. Labeled HK residues are listed in Supplementary Table 3. The relative SASA for labeled HK residues was calculated using the Rosetta RelSASA application, and crystal structures of benchmark proteins were used as input structures. A relative SASA range of 65−100% was pursued as residues within this range are very solvent-exposed.

To assess the noisiness of the exposure data, we investigated the number of false negatives in our data sets. False negatives were defined as unlabeled residues with high solvent exposure, and it has been shown that data sets can accommodate up to 35% false negative data and still meaningfully guide protein structure prediction.[26] Unlabeled STY residues with a 5−35% relative SASA and a high hydrophobic neighbor count were considered false negatives. We defined a high hydrophobic neighbor count as greater than 3.91, which is the midpoint between the average labeled hydrophobic neighbor count (4.42) and the average unlabeled hydrophobic neighbor count (3.39). False negatives within the 65−100% SASA HK residues were defined as unlabeled HK residues with greater than 80% SASA, the midpoint between the average labeled SASA and the averaged unlabeled SASA of HK residues. False positives were also calculated by assessing the number of labeled STY residues with hydrophobic neighbor counts less than 3.91 and labeled HK residues with relative SASA values less than 80%. The percentage of false negatives and false positives for each residue type set was calculated using a custom Python script.

DEPC-Guided Scoring and Model Evaluation. Based on the observed differences in Ser, Thr, and Tyr labeled and unlabeled hydrophobic neighbor counts, a score term was developed to harness these variations for structure prediction. The labeled portion of the term, STY_labeled, was calculated using eq 2:

$$\text{STY\_labeled} = \sum_{i}^{n}\left[\frac{1.0}{1+\exp(8\times(\text{hnc}_i-4.42))}-1\right] \tag{2}$$

in which n represents the number of labeled Ser, Thr, and Tyr residues, hnc_i is the hydrophobic neighbor count (see eq 1; calculated within the Rosetta score term) of the labeled Ser, Thr, or Tyr residue i, 8 is the steepness value, and 4.42 is the average hydrophobic neighbor count value calculated from the number of hydrophobic neighboring residues of labeled Ser, Thr, and Tyr residues according to the initial derivation. The per-residue STY_labeled value ranged from −1, representing agreement with the labeled residue having a hydrophobic environment, to 0, indicating disagreement because the labeled residue did not exhibit a hydrophobic environment. The unlabeled portion of the term, STY_unlabeled, was calculated as shown in eq 3:

$$\text{STY\_unlabeled} = \sum_{i}^{n}\frac{-1.0}{1+\exp(8\times(\text{hnc}_i-3.39))} \tag{3}$$

in which n is the number of unlabeled Ser, Thr, and Tyr residues, hnc_i is the hydrophobic neighbor count (calculated within Rosetta) of the particular unlabeled Ser, Thr, or Tyr residue i, 8 is the steepness value, and 3.39 is the average hydrophobic neighbor count value of unlabeled Ser, Thr, and Tyr residues in the benchmark protein crystal structures. The per-residue values also ranged from −1, indicating that the unlabeled residue had fewer hydrophobic neighbors, to 0, implying that the unlabeled residue had more hydrophobic neighboring residues and disagreed with the expected trends.

Labeled His and Lys residues were rewarded based on their relative SASA value, as shown in the HK_labeled term in eq 4:

$$\text{HK\_labeled} = \sum_{i}^{n}\left[\frac{1.0}{1+\exp(2\times(\text{relSASA}_i-0.65))}-1\right] \tag{4}$$

in which n is the number of labeled His and Lys residues, relSASA_i is the relative SASA value of the labeled His or Lys residue i, 2 is the steepness value, and 0.65 is the midpoint value of the score. The midpoint value was set as the lower end of the investigated SASA range, 65−100%, which demonstrated a measurable difference in the average relative SASA value between labeled and unlabeled residues.

Finally, the labeled and the unlabeled scores for Ser, Thr, and Tyr residues along with the portion from labeled His and Lys residues were aggregated to determine depc_ms, as shown in eq 5.

$$\text{depc\_ms} = \text{STY\_labeled} + \text{STY\_unlabeled} + \text{HK\_labeled} \tag{5}$$

The depc_ms term was used to score 10,000 ab initio models (for each of the benchmark proteins β2-microglobulin, ubiquitin, and myoglobin) and 3500 homology models (for each of the benchmark proteins carbonic anhydrase, lysozyme, and human growth hormone). The total score was calculated as a weighted superposition of the initial Rosetta score and the depc_ms score (as shown in eq 6).

$$\text{total score} = (9.0\times\text{depc\_ms score}) + \text{Rosetta Ref15 score} \tag{6}$$

A weight of 9.0 was used, similar to those reported in previous work.[27,29] A tutorial describing how to use DEPC data to predict the protein structure in Rosetta is included in the Supporting Information.

A comparison of Rosetta scoring and scoring with depc_ms was executed using several evaluation metrics. The top scoring model RMSD value was compared before and after rescoring. Additionally, the funnel-like quality, or the shape of the score versus RMSD distributions, was assessed with P_{near}. The metric P_{near} provided an insight into whether the score versus RMSD distributions featured distinctive low-energy conformations that were similar to the crystal structure. We used a funnel depth of 1.0 as proposed by Bhardwaj et al. and employed in our previous score term implementations.[27,29,46] P_{near} was calculated according to eq 7:

$$P_{\text{near}} = \frac{\sum_{m=1}^{n}\exp\left(-\frac{\text{rmsd}_m^2}{\lambda^2}\right)\exp\left(-\frac{\text{score}_m}{k_BT}\right)}{\sum_{m=1}^{n}\exp\left(-\frac{\text{score}_m}{k_BT}\right)} \tag{7}$$

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195

in which n represents the number of models generated, $score_m$ is the score of the model, and $rmsd_m$ is the RMSD of the particular model to the crystal structure. The λ value was maintained at 2.0 Å to specify which models were considered nativelike. k_BT, the effect of funnel depth, was maintained at a value of 1.0. A P_{near} value of 0 indicated no funnel-like quality, while a value of 1 signified a perfect funnel-like distribution.

■ RESULTS AND DISCUSSION

Identification of Relative SASA Ranges to Maximize Differences in Labeled and Unlabeled Residues. Based on the proposition that DEPC labeling for STY residues is sensitive to neighboring hydrophobic residues in the microenvironment,[12] we aimed to use Rosetta to elucidate a notable difference in the hydrophobic neighbor count between the labeled and unlabeled STY residues.

The six proteins in our benchmark set for which we obtained DEPC-based CL-MS data were carbonic anhydrase, ubiquitin, myoglobin, β2-microglobulin, lysozyme, and human growth hormone. We used the crystal structures of the benchmark proteins during the score development in order to identify the number of hydrophobic neighbors in the microenvironment and relative solvent exposures of the STY residues.

We identified all STY residues with a relative SASA ranging from 5 to 35%. Within the benchmark set, this encompassed 24 labeled STY residues and 22 unlabeled STY residues. This SASA range captured low-exposure STY residues, similar to those which were noted to be relevant to hydrophobic microenvironmental effects. Subsequently, we assessed the hydrophobic microenvironment for all 46 low-exposure STY residues by measuring the hydrophobic neighbor counts.

To maintain the 6 Å distance similar to DEPC molecular dimensions[12] while still accounting for neighbors likely to have a microenvironmental effect on labeling, we used a gradual neighbor count contribution method. We calculated the per-residue neighbor count by determining the contribution of neighboring hydrophobic residues based on the distance from the STY hydroxyl group. The average labeled STY hydrophobic neighbor count was determined to be 4.42, while the average unlabeled STY hydrophobic neighbor count was 3.39, as shown in Figure 1. The violin plot in Figure 1 shows the relative frequency of hydrophobic neighbor count values for labeled and unlabeled STY residues. There were less than four hydrophobic neighbors for 46% of labeled STY residues versus 68% of unlabeled STY residues while only 25% of labeled residues and 14% of unlabeled residues had a hydrophobic neighbor count greater than five. Labeled STY residues exhibited more hydrophobic neighbors than unlabeled STY residues, corroborating that STY labeling is sensitive to neighboring hydrophobic residues within the microenvironment.

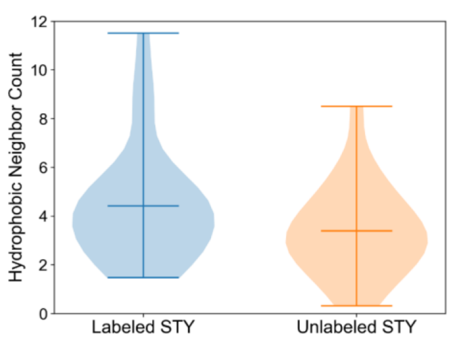


Figure 1. Violin plot demonstrating the relative frequency of different hydrophobic neighbor count values for labeled STY (blue, includes 24 residues) and unlabeled STY (orange, includes 22 residues) residues with relative SASA values of 5−35% from benchmark protein crystal structures. Mean and extrema are shown on the plot.

Using the observed trends, we developed a Rosetta score term that rewarded models containing labeled STY residues with higher hydrophobic neighbor counts and also rewarded models containing unlabeled STY residues with lower hydrophobic neighbor counts.

Additionally, since exposed His and Lys residues are more likely to be labeled by DEPC,[12] we developed a score that rewarded labeled His and Lys residues with high exposure (independent of their hydrophobic neighbor count). We used a relative SASA range of 65−100% to reward labeled His and Lys residues with high solvent exposure. The violin plot distribution for labeled and unlabeled HK relative SASA values is shown in Supplementary Figure 2, which demonstrated the expected trend that labeled HK residues were more likely to be strongly solvent-exposed. The average relative SASA for labeled residues was 0.09, higher than those of unlabeled residues. None of the unlabeled HK residues had relative SASA values greater than 0.9, while 37.5% of labeled HK residues had a relative SASA higher than 0.9.

We sought to examine the noise level of DEPC labeling data by investigating false negative data points. False negative data points within covalently labeling data sets have been previously examined; it has been suggested that 35% of exposed residues can be tolerated as false negatives while still being useful for protein structure prediction.[26] We defined false negatives in this work as unlabeled STY residues with a hydrophobic neighbor count greater than 3.91 and unlabeled HK residues with relative SASAs greater than 80%. We found that 26% of the subset of STY residues with a 5−35% SASA were false negatives and 16% of the unlabeled HK residues within a 65−100% SASA were false negatives. Both residue subsets fell well below the 35% tolerance cutoff, demonstrating that while some noise existed in our data sets, it did not impact our ability to predict accurate structures. False positive data points were defined as labeled STY residues with a hydrophobic neighbor count less than 3.91 and labeled HK residues with relative SASAs less than 80%. We determined that 15% of labeled STY residues with a 5−35% relative SASA were false positives and 16% of labeled HK residues with a 65−100% relative SASA were false positives. Subsequently, we sought to utilize the observed exposure and microenvironment trends for STY and HK residues in Rosetta protein structure prediction.

Ab initio models scored with the DEPC-guided score term showed improvement in best scoring model RMSDs and funnel-like distributions. Based on differences in hydrophobic neighbor counts for labeled and unlabeled STY residues (Figure 1) and differences in the SASA for labeled and unlabeled HK residues (Supplementary Figure 2), we proceeded to develop a score term that rewarded the desired trends. An overview of the score term is shown in Figure 2. We mapped the label status of labeled and unlabeled residues onto protein models using the DEPC-based CL-MS data. Figure 2a depicts the inputs of the score term, which included labeling data as the label status (L for labeled and U for unlabeled) and appropriate residue number along with ab initio or homology protein models. Additional details can be found in the tutorials in Supplementary Note 1. We calculated the relative SASA for

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195

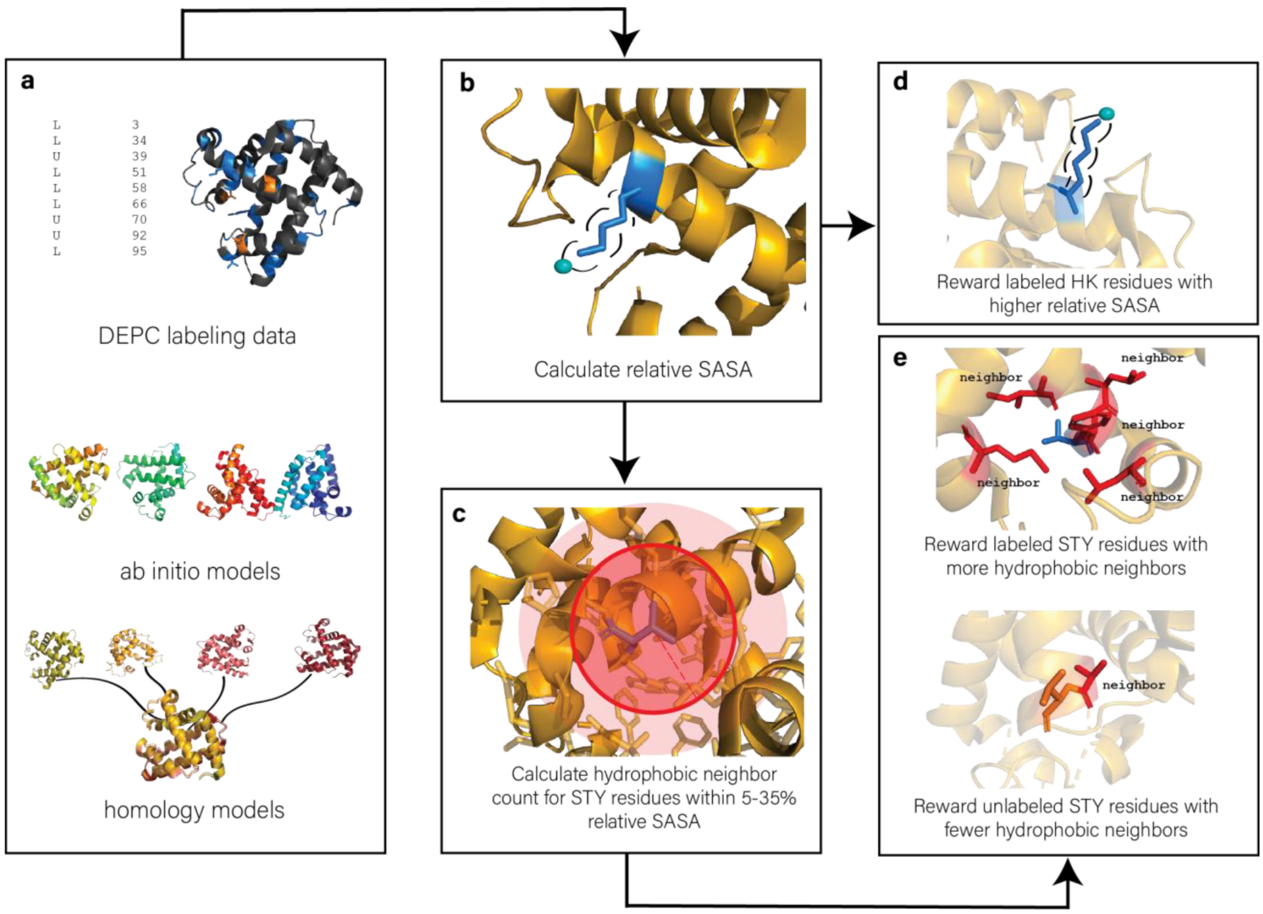


Figure 2. Overview of the DEPC score term (depc_ms) algorithm. The DEPC score term required CL-MS labeling data (residue numbers and label status) as the input, along with input structures, which were either generated with homology or ab initio modeling (a). The relative SASA was calculated for all residues listed in the input file (b). If the residue was STY, additionally the hydrophobic neighbor count was calculated for residues with relative SASAs between 5 and 35% (c). Labeled HK residues with higher relative SASAs were rewarded (d). Labeled STY residues with more hydrophobic neighbors and unlabeled residues with less hydrophobic neighbors were rewarded as well (e).

all mapped residues (Figure 2b) and hydrophobic neighbor counts for buried STY residues (Figure 2c). The score term included components that rewarded labeled STY residues with high numbers of hydrophobic neighbors (Figure 2e), rewarded unlabeled STY residues with low numbers of hydrophobic neighbors (Figure 2e), and rewarded labeled HK residues with high solvent exposure (Figure 2d). While our initial analysis of hydrophobic neighbor counts and relative SASA ranges relied on crystal structures, no crystal structures were used in model generation or score term evaluation. To test the DEPC-guided scoring, we generated 10,000 ab initio models for each protein within our benchmark set. Upon examination of the ab initio models, we noticed that three of the benchmark protein model sets did not contain any models under a 5 Å RMSD to the crystal structure. The DEPC score was designed to distinguish nativelike models (RMSD < 5 Å) from incorrect models (RMSD > 10 Å). In order to have higher quality models present in all of the benchmark cases, homology models were generated for these three benchmark proteins. The results of scoring these homology models will be discussed in the next section.

For the three proteins whose ab initio model distributions included nativelike models (ubiquitin, myoglobin, and β2-microglobulin), we tested our score term, depc_ms, by adding the DEPC score to the total Rosetta score. As seen in Figure 3, the best scoring model RMSD values improved from Rosetta scoring (Figure 3a) to scoring with Rosetta and depc_ms, our DEPC-guided score term (Figure 3b). The RMSD of the best scoring model for β2-microglobulin improved from an RMSD of 3.14 to 2.13 Å, while ubiquitin improved from an RMSD of 3.16 to 1.97 Å. Myoglobin saw a notable improvement from an RMSD of 7.11 to 1.36 Å when scoring with the depc_ms term.

Additionally, the funnel-like quality of the distributions was quantified by the P_{near} value, with a higher P_{near} value (near 1) indicating more funnel-like quality and a lower P_{near} value (near 0) indicating lack of any funnel-like quality. We noticed that all distributions became more funnel-like, i.e., an increased P_{near} value, with DEPC labeling data included in scoring, indicating that we were selecting lower-energy conformations that were more similar to the native. For β2-microglobulin, P_{near} values increased from 0.22 with Rosetta to 0.31 with depc_ms; ubiquitin improved from 0.08 to 0.32. Myoglobin exhibited the largest improvement in the P_{near} value, increasing from 0.07 to 0.33 with depc_ms scoring.

Homology Model Predictions Also Improved upon Scoring with DEPC Data. For carbonic anhydrase, human growth hormone, and lysozyme, the best model generated with ab initio modeling had an RMSD value greater than 5 Å to the crystal structure. We thus sought to generate additional models with Rosetta's comparative modeling protocol. The homology modeling templates with their respective sequence identities and similarities are shown in Supplementary Table 1. By

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195

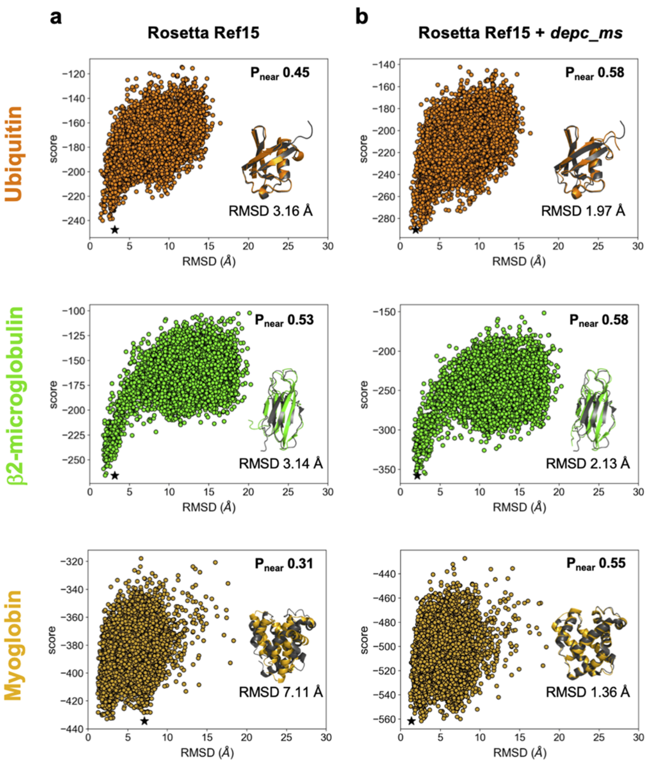


Figure 3. Score versus the RMSD to the crystal structure for 10,000 ab initio models for (a) Rosetta without DEPC labeling data and (b) Rosetta with DEPC labeling data. Best scoring models are marked by a black star and shown in color aligned to the crystal structure (gray). P_{near} values are listed.

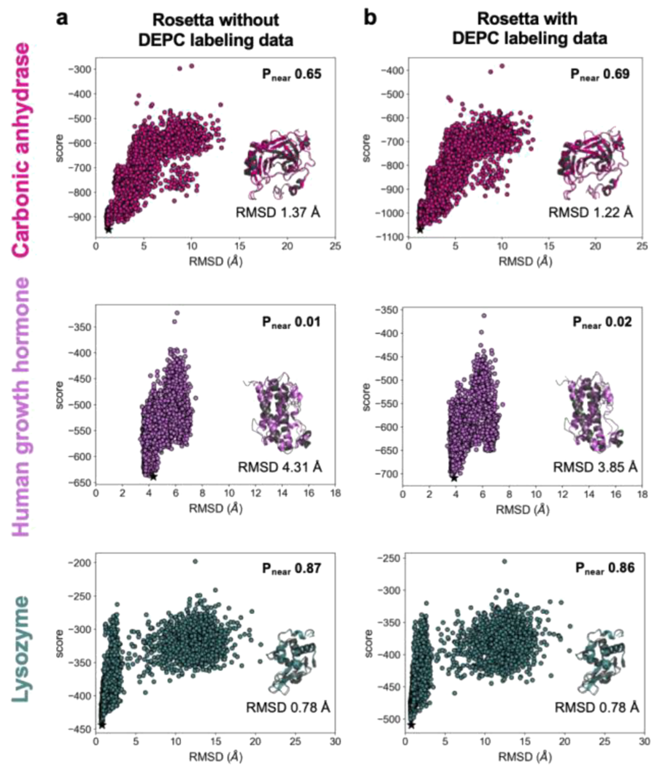


Figure 4. Score versus the RMSD to the crystal structure for 3500 homology models for (a) Rosetta without DEPC labeling data and (b) Rosetta with DEPC labeling data. Best scoring models are marked by a black star and shown in color aligned to the crystal structure (gray). P_{near} values are listed.

generating 500 models per template and using multiple templates per protein, we were able to generate a distribution of models with varying RMSD values. We scored all models with depc_ms and subsequently added the score to the Rosetta score. The total score versus RMSD plots along with the best scoring model aligned with the crystal structure are shown in Figure 4. While the models identified by Rosetta were already significantly better for homology models (as compared to the ab initio models in the last section), scoring with DEPC data further improved model selection consistently. The human growth hormone best scoring model RMSD improved from 4.31 Å with Rosetta without labeling data to 3.85 Å with Rosetta with DEPC labeling data by way of depc_ms. The lysozyme best scoring model RMSD (0.78 Å) stayed constant from Rosetta to scoring with DEPC data at already accurate atomic detail. Finally, the carbonic anhydrase best scoring model RMSD improved from 1.33 to 1.22 Å.

Improvements in the funnel-like quality of the distributions, P_{near}, were also noted for both human growth hormone and carbonic anhydrase. The P_{near} value of lysozyme for which the best scoring model had a sub-angstrom RMSD, stayed at 0.83, an already near-perfect value. Human growth hormone P_{near} slightly improved from 0.01 to 0.02, while carbonic anhydrase P_{near} improved from 0.65 to 0.69 with DEPC scoring. Overall, our scoring methodology with DEPC labeling data successfully improved the best scoring model quality and distribution funnel-like quality.

■ CONCLUSIONS

To employ DEPC labeling MS data in protein structure prediction, we analyzed the difference in hydrophobic neighbor counts between labeled and unlabeled STY residues and used labeled HK residues with high solvent exposure. Our benchmark set, consisting of DEPC-labeled ubiquitin, *β*2-microglobulin, myoglobin, human growth hormone, carbonic anhydrase, and lysozyme, was used to explore the utility of DEPC labeling in protein structure elucidation. We developed a novel Rosetta score term which rewarded STY residues known to be DEPC-labeled if those residues exhibited a hydrophobic microenvironment and rewarded unlabeled STY residues that lacked such hydrophobic microenvironment. Additionally, the term rewarded labeled HK residues with high solvent exposure. In a test of our algorithm, we noted that usage of DEPC data improved the best scoring model RMSD and the funnel-like quality of the model distribution. For the six benchmark proteins, we saw improvement in the prediction quality for both ab initio and homology models. Notably, we elucidated accurate atomic detail for all six proteins upon employment of DEPC labeling data. The advantageous qualities of the DEPC label, such as single product generation and ease of commercial availability, along with our DEPC-guided Rosetta modeling that is solely based on the label status and computationally determined exposure metrics underscore the huge potential of DEPC labeling for protein structure determination.

While our work was primarily focused on DEPC labeling, different modeling strategies for other types of labels have previously been developed. The nature of the label generally dictates the modeling strategies warranted. For instance, hydroxyl radical protein footprinting is sufficiently modeled with solvent exposure alone.[29] Modeling HDX labeling benefits from accounting for both residue exposure and

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195

flexibility.[47] Other labels with microenvironmental effects would benefit from further analysis regarding modeling strategies to employ. Future work will continue to pursue covalent labeling data implementation into model generation protocols. Additionally, we aim to test this methodology on larger (500−1000 residues) proteins. We plan to examine the accuracy of our scoring function when utilizing DEPC data as the labeling extent. Further studies will also emphasize the role of dynamics and microenvironmental effects in covalent labeling.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c00395.

> DEPC-modified residue structures, violin plot comparing the relative SASA between labeled and unlabeled HK residues, table of templates used for homology model generation, labeling data for STY and HK residues, and step-by-step command line tutorial for homology model generation and score term usage (PDF)
>
> Tutorial files including 500 ab initio models, input files, and homology modeling script and weights (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; Email: lindert.1@osu.edu

Authors

Sarah E. Biehn − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States*

Patanachai Limpikirati − *Department of Food and Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok 10330, Thailand*

Richard W. Vachet − *Department of Chemistry, University of Massachusetts, Amherst, Massachusetts 01003, United States;* orcid.org/0000-0003-4514-0210

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c00395

Author Contributions

P.L. and R.V.W. performed DEPC labeling experiments. S.E.B. and S.L. designed the Rosetta prediction algorithm and performed benchmark analysis. S.E.B., R.V.W., 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 and Vachet labs for useful discussions and the Ohio Supercomputer Center for computational resources.[48] Research reported in this publication was supported by NSF (CHE 1750666) to S.L. and by the NIH (R01 GM075092) to R.W.V.

■ REFERENCES

(1) Konermann, L.; Pan, J.; Liu, Y.-H. *Chem. Soc. Rev.* **2011**, *40*, 1224−1234.
(2) Sinz, A. *Mass Spectrom. Rev.* **2006**, *25*, 663−682.
(3) Liu, X. R.; Zhang, M. M.; Gross, M. L. *Chem. Rev.* **2020**, *120*, 4355−4454.
(4) Xu, G.; Chance, M. R. *Chem. Rev.* **2007**, *107*, 3514−3543.
(5) Cheng, M.; Zhang, B.; Cui, W.; Gross, M. L. *Angew. Chem. Int. Ed.* **2017**, *56*, 14007−14010.
(6) Jumper, C. C.; Schriemer, D. C. *Anal. Chem.* **2011**, *83*, 2913−2920.
(7) Limpikirati, P.; Liu, T.; Vachet, R. W. *Methods* **2018**, *144*, 79−93.
(8) Glocker, M.; Kalkum, M.; Yamamoto, R.; Schreurs, J. *Biochemistry* **1996**, *35*, 14625−14633.
(9) Dage, J. L.; Sun, H.; Halsall, H. B. *Anal. Biochem.* **1998**, *257*, 176−185.
(10) Mendoza, V. L.; Vachet, R. W. *Mass Spectrom. Rev.* **2009**, *28*, 785−815.
(11) Mendoza, V. L.; Vachet, R. W. *Anal. Chem.* **2008**, *80*, 2895−2904.
(12) Limpikirati, P.; Pan, X.; Vachet, R. W. *Anal. Chem.* **2019**, *91*, 8516−8523.
(13) Zhou, Y.; Vachet, R. W. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 899−907.
(14) Zhou, Y.; Vachet, R. W. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 708−717.
(15) Xie, B.; Sood, A.; Woods, R. J.; Sharp, J. S. *Sci. Rep.* **2017**, *7*, 4552.
(16) Benesch, J. L.; Ruotolo, B. T. *Curr. Opin. Struct. Biol.* **2011**, *21*, 641−649.
(17) Jia, R.; Martens, C.; Shekhar, M.; Pant, S.; Pellowe, G. A.; Lau, A. M.; Findlay, H. E.; Harris, N. J.; Tajkhorshid, E.; Booth, P. J. *Nat. Commun.* **2020**, *11*, 6162.
(18) Murcia Rios, A.; Vahidi, S.; Dunn, S. D.; Konermann, L. *J. Am. Chem. Soc.* **2018**, *140*, 14860−14869.
(19) Hall, Z.; Politis, A.; Robinson, C. V. *Structure* **2012**, *20*, 1596−1609.
(20) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. *Structure* **2015**, *23*, 791−799.
(21) Eschweiler, J. D.; Farrugia, M. A.; Dixit, S. M.; Hausinger, R. P.; Ruotolo, B. T. *Structure* **2018**, *26*, 599−606.e3.
(22) Mistarz, U. H.; Chandler, S. A.; Brown, J. M.; Benesch, J. L.; Rand, K. D. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 45−57.
(23) Busch, F.; VanAernum, Z. L.; Ju, Y.; Yan, J.; Gilbert, J. D.; Quintyn, R. S.; Bern, M.; Wysocki, V. H. *Anal. Chem.* **2018**, *90*, 12796−12801.
(24) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P. *Nat. Methods* **2020**, *17*, 665−680.
(25) 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. *J. Chem. Theory Comput.* **2017**, *13*, 3031−3048.
(26) Aprahamian, M. L.; Lindert, S. *J. Chem. Theory Comput.* **2019**, *15*, 3410−3424.
(27) Aprahamian, M. L.; Chea, E. E.; Jones, L. M.; Lindert, S. *Anal. Chem.* **2018**, *90*, 7721−7729.
(28) Kahraman, A.; Malmström, L.; Aebersold, R. *Bioinformatics* **2011**, *27*, 2163−2164.
(29) Biehn, S. E.; Lindert, S. *Nat. Commun.* **2021**, *12*, 341.
(30) 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. *Proc. Natl. Acad. Sci.* **2019**, *116*, 8143−8148.
(31) Seffernick, J. T.; Harvey, S. R.; Wysocki, V. H.; Lindert, S. *ACS Cent. Sci.* **2019**, *5*, 1330−1341.
(32) Hartlmüller, C.; Göbl, C.; Madl, T. *Angew. Chem. Int. Ed.* **2016**, *55*, 11970−11974.
(33) Seffernick, J. T.; Lindert, S. *J. Chem. Phys.* **2020**, *153*, 240901.

(34) Leelananda, S. P.; Lindert, S. *J. Chem. Inf. Model.* **2019**, *60*, 2522−2532.
(35) Song, Y.; DiMaio, F.; Wang, R. Y.-R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D. *Structure* **2013**, *21*, 1735−1742.
(36) Miles, E. W. *Methods Enzymol.* **1977**, *47*, 431−442.
(37) Ziemianowicz, D. S.; Sarpe, V.; Schriemer, D. C. *Anal. Chem.* **2019**, *91*, 8492−8499.
(38) Barth, M.; Bender, J.; Kundlacz, T.; Schmidt, C. *J. Proteomics* **2020**, *222*, No. 103793.
(39) Liu, T.; Limpikirati, P.; Vachet, R. W. *Anal. Chem.* **2019**, *91*, 15248−15254.
(40) Pan, X.; Limpikirati, P.; Chen, H.; Liu, T.; Vachet, R. W. *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 658−665.
(41) Liu, T.; Marcinko, T. M.; Vachet, R. W. *J. Am. Soc. Mass Spectrom.* **2020**, 1544.
(42) Borotto, N. B.; Zhou, Y.; Hollingsworth, S. R.; Hale, J. E.; Graban, E. M.; Vaughan, R. C.; Vachet, R. W. *Anal. Chem.* **2015**, *87*, 10627−10634.
(43) Limpikirati, P.; Hale, J. E.; Hazelbaker, M.; Huang, Y.; Jia, Z.; Yazdani, M.; Graban, E. M.; Vaughan, R. C.; Vachet, R. W. *MAbs* **2019**, *11*, 463−476.
(44) Kim, D. E.; Chivian, D.; Baker, D. *Nucleic Acids Res.* **2004**, *32*, W526−W531.
(45) Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. *Protein Sci.* **2014**, *23*, 47−55.
(46) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V.; Kaas, Q.; Eletsky, A. *Nature* **2016**, *538*, 329−335.
(47) Marzolf, D. R.; Seffernick, J. T.; Lindert, S. *J. Chem. Theory Comput.* **2021**, *17*, 2619−2629.
(48) *Ohio Supercomputer Center*. 1987. Ohio Supercomputer Center: Columbus, OH http://osc.edu/ark:/19495/f5s1ph73.

https://doi.org/10.1021/acs.analchem.1c00395
Anal. Chem. 2021, 93, 8188−8195