analytical chemistry

pubs.acs.org/ac Article

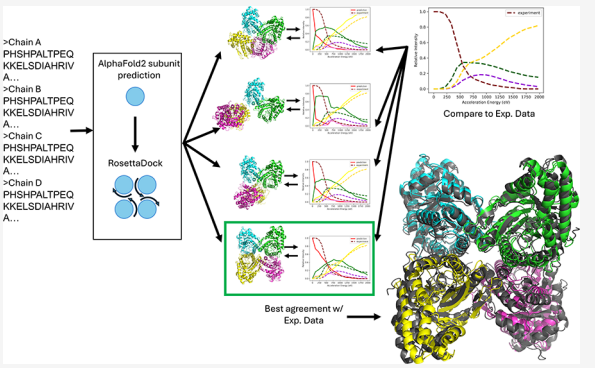
Energy Resolved Mass Spectrometry Data from Surfaced Induced Dissociation Improves Prediction of Protein Complex Structure

Robert M. Bolz, Justin T. Seffernick, Zachary C. Drake, Sophie R. Harvey, Vicki H. Wysocki, and Steffen Lindert*

Cite This: *Anal. Chem.* 2025, 97, 2375−2383 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Native Mass Spectrometry (nMS) is a versatile technique for elucidating protein structure. Surface-Induced Dissociation (SID) is an activation method in tandem MS predominantly employed for determining protein complex stoichiometry alongside information about interface strengths. SID-nMS data can be collected over a range of acceleration energies, yielding Energy Resolved Mass Spectrometry (ERMS) data. Previous work demonstrated that the onset and appearance energy from SID-nMS can be used in integrative computational and experimental modeling to guide multimeric structure determination in some cases. However, the appearance energy is a single data point, while the ERMS data provide a full pattern of interface breakage. We hypothesized that incorporation of ERMS data into multimeric protein structure prediction would significantly outperform appearance energy. To test this hypothesis, we generated models of 20 protein complexes with RosettaDock using subunits generated from AlphaFold2. We simulated the ERMS data for each predicted model and rescored based on its agreement to experimental ERMS data. We demonstrated that more accurately predicted models exhibited simulated ERMS data in better agreement with the experimental data. As part of our ERMS-based rescoring, we matched or improved the RMSD of the best scoring model compared to Rosetta in 16 out of 20 cases, with 4 out of 20 cases improving to become a highly accurate (below 5 Å) structure. Finally, we benchmarked our method against our previously published appearance energy-based rescoring and showed improvement in 14 out of 20 cases, with 6 out of 20 becoming a highly accurate (below 5 Å) model. Our method is freely available through Rosetta Commons, with a usage tutorial and test files provided in the Supporting Information.



INTRODUCTION

Alzheimer's,[1] breast cancer,[2] and Parkinson's[3] are diseases resulting from aberrant protein interactions. Furthermore, 86% of proteins are found to form a complex in vivo,[4] providing an impetus to better understand the complexes that proteins form. Conventional structure determination methods such as X-ray Crystallography,[5] Nuclear Magnetic Resonance,[6] and cryogenic Electron Microscopy[7] serve as the current gold standards in structure determination. These techniques all offer the ability to obtain an atomic or near-atomic resolution structure. Each offers unique strengths and weaknesses for determining a complex structure. However, throughput and sample preparation requirements[8] for these techniques limit their ability to resolve all biological systems. Native Mass Spectrometry[9] is an alternative method to traditional protein structure determination. While not able to provide a complete three-dimensional protein structure, nMS is fast, has less stringent sample requirements, and is able to probe structure for systems normally inaccessible to high-resolution techniques like the ones listed above.[9] For this reason, nMS is a popular tool to collect structural information about many biological systems.

Native Mass Spectrometry, using one or more of a variety of ion activation methods, can elucidate protein structural elements.[9] In SID-nMS, a protein complex is accelerated toward a collision surface.[8] The energy transferred from kinetic energy upon the collision with the collision surface is converted to internal, vibrational energy. This has the potential to break the noncovalent interfaces between the subunits of the complex. In addition to yielding data on complex stoichiometry, SID-nMS provides information on the interface size and strength of a protein complex, with the weakest interfaces cleaving at the lowest energies. By accelerating the complex toward the surface at different energies and quantifying the resulting subcomplexes, a characteristic dissociation curve for each complex can be obtained. This technique is particularly valuable because, in

Received: October 30, 2024
Revised: January 11, 2025
Accepted: January 15, 2025
Published: January 24, 2025

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

many cases, it does not initially denature the individual subunits of a complex, allowing for near-native interface information to be extracted.[8] Because of the short time scale of the nMS spectral collection, the kinetically trapped complexes remain structured in the compact native solution state as they travel through the gas phase. The collision surface is rigid and resists charge transfer, allowing for the conversion of kinetic (collision) energy to internal vibrational energy. When collected over a range of acceleration energies, the relative abundance of precursor and subcomplexes collected with SID-nMS (or other activation method) is referred to as Energy-Resolved Mass Spectrometry data or ERMS data. This ERMS data can be visualized showing the relative abundance of precursor and fragments plotted against the acceleration energy at which the abundance was measured, referred to as an ERMS plot. Other activation methods such as collision induced dissociation (CID) can provide quaternary information on protein structure. Collected over a range of energies these data can be transformed into ERMS data, but CID typically produces monomer and $(n-1)$-mer fragments, providing information on stoichiometry rather than diagnostic substructure fragments.[10] As an example, SID of a dimer-of-dimers structure will produce dimeric products at the lower SID energies, and SID of a dimer-of-trimers would produce trimeric products. CID for a dimer-of-dimers tetramer typically produces monomer and trimer, confirming a tetrameric complex (but does not indicate the arrangement of subunits), while CID of a dimer-of-trimers typically produces a monomer and pentamer, confirming a hexameric structure (but with no confirmation of substructure). One of our earlier publications[11] includes representative CID spectra for many of the complexes under consideration here and CID ERMS can be found in Quintyn et al. for three tetrameric complexes.[10] In previous work,[11] we assigned the appearance energy (AE) for SID protein complex dissociation as the acceleration energy at which 10% of the precursor protein complex had dissociated into a given set of subcomplex fragments (see Figure S1 for a representative ERMS plot showing the definition of AE).[12] Notably, every AE only constitutes a single data point from within the full information-rich ERMS data.[10]

All structural data collected from MS are low-resolution, meaning that there is not enough information to directly determine the atomic-level structure. One method to allow for a full structure to be predicted from MS data is integrative modeling, i.e., coupling experimental data with computational structure prediction. This method has been used in the past to elucidate protein structure from a variety of MS data.[13,14,11,15−35] Specifically, we have used SID-nMS data for protein structure interpretation or prediction.[11,13,14] First, we showed that we can accurately predict SID-AE values for a large number of protein complexes from the properties of the complex interface.[11] Comparing computed AE values for a range of predicted structural conformations of a protein complex with experimentally determined AEs provided a measure of the structural agreement of the predicted complex with the native protein structure. Incorporating this measure into a Rosetta[36] score term allowed us to improve accuracy in protein complex prediction.[13] This method was subsequently improved through the use of more advanced docking algorithms and the addition of simultaneous cryo-EM density modeling using low resolution maps.[14] Collectively, this work demonstrated the viability of using SID-AE data in protein structure prediction. However, we speculated that using single-point AEs exclusively underutilized the information encoded in the SID-nMS experiment. We hypothesized that using the full multipoint ERMS data to predict protein structure may be more beneficial.

As a first step in using ERMS to predict structure, we showed that it is possible to predict SID-ERMS from interface properties of a given protein complex.[37] We first measured the interface size and strength using Rosetta.[38] We then developed an ERMS prediction algorithm that simulates breaking of the interface to determine the characteristic fragmentation pattern. This simulated fragmentation was repeated for all interfaces of the complex as well as over a range of acceleration energies to produce fully simulated ERMS data.

In 2020, the method AlphaFold2,[39] developed by DeepMind, revolutionized the world of computational protein structure prediction. Monomer prediction using AlphaFold2 has been shown to be on average 76.7% accurate (measured through mean GDT_TS).[40] This makes it the current leading tool in protein structure prediction. However, its multimer prediction is still lagging. Independent benchmarking of AlphaFold-Multimer shows an average of 50% accuracy (measured through DockQ) of complex models, with that average trending lower for larger complexes.[41] Consequently, sparse experimental data are still beneficial to ensure a more accurate protein complex prediction. While Rosetta is ideally suited for incorporation of experimental data, it is not easily possible to incorporate experimental data into AlphaFold-Multimer due to its lack of explicit score functions. Hence, we hypothesized that the coupling of AlphaFold2 monomer prediction with SID-ERMS-guided RosettaDock complex modeling was an optimal transitional strategy for integrative SID-nMS modeling and would leverage the individual strengths of the three methods.

In this work, we used protein−protein docking to predict complexes for 20 systems from the SID-ERMS data. We demonstrated a correlation between the accuracy of the predicted structure and the Root Mean Square Error (RMSE) of the predicted and experimental SID-ERMS data. We then rescored these docked structures based on their deviation between predicted and experimental SID-ERMS. We matched the accuracy of Rosetta or saw improvement in Root Mean Square Deviation (RMSD) of the best scoring model in 16/20 test cases, with 4/20 cases improving to below 5 Å RMSD. Finally, we showed that using the full SID-ERMS data for structure prediction yielded superior results as compared to our previous predictions that used only SID-AE. This method is freely available through Rosetta Commons with a tutorial and test files included in the Supporting Information.

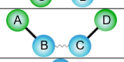
METHODS

Benchmark Data Set. The benchmark set was comprised of proteins for which experimental SID-ERMS data has been previously published.[11] This set contained six homodimers with C2 symmetry, eight homodimers with D2 symmetry, three homopentamers with C5 symmetry, one homohexamer with C6 symmetry, and two heterotetramers with C2 and D2 symmetry, respectively. The names, PDB IDs, complex types, connectivities, and symmetries for each protein can be found in Table 1.

Subunit Generation. To obtain protein complex structural models, we generated structures using ensemble docking in Rosetta. We pursued two docking strategies: 1) docking of subunits that were derived from the deposited PDB structure and 2) docking of subunits that were generated through Alphafold2. Subunits derived from the PDB structure were created by extracting the subunit from the original PDB. The subunits created by Alphafold2 were generated using the 2.2.2

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

Table 1. The 20 Proteins Comprising Our Benchmark Dataset[a]

Name	PDB ID	Complex Type	Connectivity	Symmetry
Enolase	1E9I	homodimer		C2
DeoC	1KTN	homodimer		
IspD	1VGT	homodimer		
Lysozyme	4R0F	homodimer		
Beta-lactoglobulin	6QI6	homodimer		
Triose phosphate isomerase	8TIM	homodimer		
Pyruvate kinase	1AQF	homotetramer		D2
Avidin (neutravidin)	1AVE	homotetramer		
Concanavalin A	1JBC	homotetramer		
Streptavidin	1SWB	homotetramer		
Carbonic anhydrase	1T75	homotetramer		
Uracil phosphoribosyl transferase	2EHJ	homotetramer		
Transthyretin	5HJG	homotetramer		
D-sialic acid aldolase	6ALD	homotetramer		
Hemoglobin	1GZX	heterotetramer		Pseudo-D2
Tryptophan synthase	1WBJ	heterotetramer		C2
Cholera toxin B	1FGB	homopentamer		C5
C-reactive protein	1GNH	homopentamer		
Serum amyloid P	1SAC	homopentamer		
Hfq	1HK9	homohexamer		C6

[a]Information is presented on the name, PDB ID, complex type, connectivity, and symmetry for each protein complex.

release of Alphafold2. Full multiple sequence alignments were generated for each prediction by using the flag−use_precomputed_msas = False. Full sequence databases were used with the−db_preset = full_dbs flag. The date of subunit generation was set as the "max template date" (max_template_date = 2023−11−01). Subunits of nontetramer complexes were predicted as monomers. Because of the nonredundancy of the interfaces, tetramer subunits were predicted as homodimers or heterodimers using AlphaFold Multimer, using the same flags. A full explanation of the considerations for D2 and C2 tetramer interfaces is presented in the Ensemble Docking section of the methods. An explanation for each generated subunit is provided as part of Table S1.

The individual subunits were perturbed using Rosetta relax, backrub, and normal-mode analysis (nma), respectively, to remove bias away from the starting structure and also to sample flexibility. 10 perturbed structures were generated, for each method, yielding 30 total subunits for each of the 20 protein complexes.

Ensemble Docking. The 30 perturbed subunits were then prepacked. Prepacking performs side chain optimization for docking partners, resulting in more accurate interface interactions. The prepacked subunits were ensemble docked using RosettaDock, docking A and B subunits in the case of the nontetramers, and AB and CD subunits for the tetramers. These subunits were paired and allowed to randomly orient by using the -randomize2 flag. For each system, a total of 10,000 complex models were generated in this fashion. The structures were subsequently relaxed using Rosetta relax.

Full complexes were generated for the homodimers. For C5 pentamers (and C6 hexamers), the interface between subunit A_B is essentially the same as B_C, C_D, D_E, and E_A (or E_F and F_A for the hexamer). For this reason, the dimer subcomplex form of these homomers is sufficient to simulate the full ERMS data. Therefore, we generated only homodimer subcomplexes for homopentamers (C5 symmetry) and homohexamers (C6 symmetry). Homotetramers with D2 and C2 symmetry have three different interfaces that need to be modeled for these complexes, with interfaces A_B equal to C_D, A_D equal to B_C, and A_C equal to B_D. Heterotetramers with D2 symmetry have four different interfaces that need to be modeled for these complexes, with interfaces A_B equal to C_D, and A_C equal to B_D, but unique interfaces A_D and B_C. Likewise, heterotetramers with C2 symmetry have equal interfaces A_B and C_D, and a unique interface B_C. Therefore, these complexes were ensemble docked as the full complex (AB_CD), in order for us to be able to simulate the full ERMS data.

ERMS Prediction. For each of the 20 protein systems, each of the 10,000 relaxed complexes had their ERMS predicted using the SID_ERMS_prediction application in Rosetta.[37] The ERMS prediction itself assumed that no secondary fragmentation of the broken complex occurs. It is likely, however, that secondary fragmentation does occur at higher energies. This is seen in experimental spectra by a shift in the products (e.g., with high energy input into the precursor complex, monomers and lower order multimers may increase in relative abundance as secondary fragments of intermediate multimers; a trimer could further fragment to monomer and dimer). A recent publication[42] shows that preactivation prior to SID may also increase the extent of secondary fragmentation, as expected. In the ERMS data used in the present paper, secondary fragmentation products likely overlap with primary products in some cases (e.g., monomers may be a primary product across part of the energy range), but then a change in ERMS slope and a steeper increase in monomer corresponding to a decrease in another fragment (e.g., a dimer decrease that tracks monomer increase) suggests secondary fragmentation. This has not yet been incorporated into our approach and will be the focus of future work (see the "Conclusions" section). The predicted ERMS was compared to the experimental ERMS[11,37,43] for that complex using RMSE, with equal weighting given to all energies for each curve.

Rescoring. We developed a score term for Rosetta based on agreement between the experimental ERMS data and the simulated data of the prediction. Each model was scored first using the Rosetta ref2015[44] score function. A scalar RMSE score was then added to the ref2015 score (eq 1) with a weight of 1000 (eq 2). The RMSE score term is a sigmoidal function with an amplitude of 1 (eq 3). The amplitude of the function is the maximum allowable penalty score for the model, providing a ceiling for the score term (Figure 1). The scalar 0.1 present in the denominator influences the steepness of the inflection of the score term (eq 3). A floor (through a conditional statement, eq 4) is additionally set for this function such that if the value of the f(RMSE) term is lower than zero, then the $\text{Score}_{\text{RMSE}}$ term will instead be zero (Figure 1). Multiplying the RMSE to scale the value allows the input RMSE to be the primary determinant of the penalty for the model (eq 3). Finally, to center the inflection point of the RMSE score term approximately within the range of observed RMSEs, the midpoint of the RMSE range weighted by 0.9 (across the 10,000 models) was incorporated into the exponent of the denominator. The midpoint is the geometric center of the maximum and minimum RMSE values (eq 3).

$$\text{SIDScore} = \text{Score}_{\text{ref2015}} + \text{weight}_{\text{RMSE}} \cdot \text{Score}_{\text{RMSE}} \quad (1)$$

$$\text{weight}_{\text{RMSE}} = 1000 \quad (2)$$

$$f(RMSE) = 1 - \frac{1}{(0.1 + e^{50 \cdot (RMSE - 0.9 \cdot mid_{RMSE})})} \quad (3)$$

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

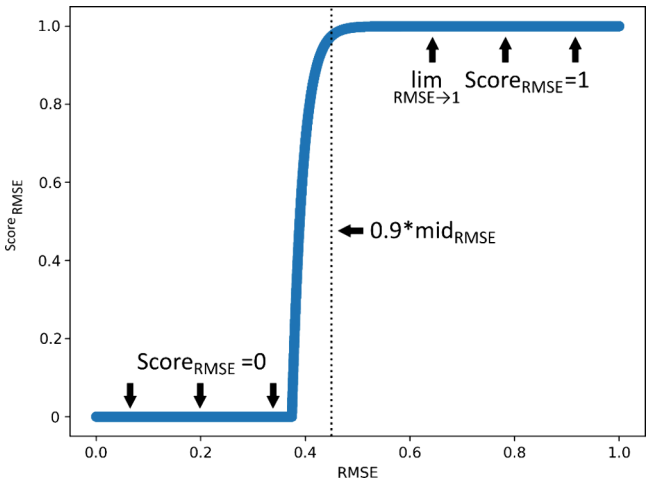


Figure 1. Representative $Score_{RMSE}$ term for a uniform distribution of RMSE values ranging from 0 to 1. The upward pointing arrows show the point at which the function has asymptotically approached one, labeled $Score_{RMSE} = 1$. The downward facing arrows show the point at which the conditional expression has been applied, setting the value of the function to zero, labeled $Score_{RMSE} = 0$. The black dotted line shows the geometric center of the RMSE data multiplied by 0.9, labeled $0.9*mid_{RMSE}$.

$$Score_{RMSE} = \begin{cases} Score_{RMSE} = 0 & f(RMSE) < 0 \\ Score_{RMSE} = f(RMSE) & f(RMSE) \geq 0 \end{cases} \quad (4)$$

RMSD Measurement. Root Mean Square Deviations (RMSDs) were measured using PyMOL.[45] The Cα RMSD with no outlier rejection was measured for each ensemble docked model against the reference structure.

RESULTS AND DISCUSSION

AlphaFold2 Monomer Prediction Produced Accurate Subunit Structures. Before generating the protein complex models, we first evaluated the accuracy of the AlphaFold2-derived subunits. The subunits were highly accurate, with 19/20 subunits having RMSD values below 2 Å. The one exception to this was the homotetramer pyruvate kinase (PDB: 1AQF), with an RMSD of 3.69 Å. This incorrect prediction of the subunit affected the assembly of the complex, which will be explained further in the subsequent results. RMSDs of each generated subunit are shown in Table S1.

Accurately Predicted Protein Structures Showed Agreement with Experimental ERMS Data. We hypothesized that SID-ERMS data contain structural information about protein complexes. To test this, we first sought to show a relationship between the accuracy of the simulated ERMS data and the accuracy of the protein structure used for the simulation. For each of the 20 proteins in our data set, we generated 10,000 ensemble docked structures, starting from the PDB-derived subunits and the Alphafold2-generated subunits (400,000 structures total), respectively. This provided a data set of structures with a wide range of accuracies (i.e., RMSDs to the native structure). The ERMS data was simulated for each of the 400,000 predicted protein complexes. Figure 2A shows the comparison of the RMSE of the simulated ERMS data to the RMSD of the 10,000 predicted complexes for Tryptophan synthase (PDB: 1WBJ) for the AlphaFold2 data set. The corresponding plots for the remainder of the benchmark set can be found in Figure S2 and Figure S3 for AlphaFold2- and PDB-derived subunits, respectively. The results were similar regardless of the input structure. In our analysis, we focused on examining the results from the complexes of Alphafold2-derived subunits because they more accurately reflected a real-world modeling case for a completely unknown structure, though the performance of the PDB structure complexes was largely the same if not slightly better.

Next, we sought to examine the correlation between the RMSD and RMSE across the entire benchmark set. Figure 2B shows the correlation between RMSE and RMSD with binned RMSD values for tryptophan synthetase, with error bars. See Figure S4 and Figure S5 for the corresponding plots for the rest of the set. The Alphafold2 plots for pyruvate kinase (PDB: 1AQF), lacked a strong relationship between RMSD and RMSE because of the incorrect AlphaFold2 subunit prediction (see Figure S4), highlighting the importance of having an accurate subunit structure. The general trends observed in the RMSE/RMSD plots were roughly the same for complexes derived from the docking of Alphafold2 subunits and PDB structure subunits. In general, we observed that bins corresponding to higher RMSD values also exhibited higher average RMSE values. To quantify this observation, we investigated the relative change in the RMSE from each bin to the first bin. Figure 2C shows this relative change for tryptophan synthase, whereas Figure 2D shows the relative changes in RMSE for the entire benchmark set with the red trendline denoting the average across all proteins. In this plot, we observed a positive average trendline over the range of RMSDs from our benchmark set. This confirmed our hypothesis that higher RMSD values generally correlated with increased disagreement to the experimental data, suggesting that ERMS data encoded viable structural information about protein complexes. The equivalent plot for complexes from the docking of PDB structure subunits can be found in Figure S6, showing the same trend. We subsequently sought to utilize this correlation in protein structure prediction.

Rescoring with ERMS Data Improved Protein Complex Prediction Compared to Rosetta Prediction. Our analysis in the previous section further demonstrated the presence of structural information within SID-ERMS data, as was seen in previous studies,[11,14] by linking the accuracy of simulated ERMS data and the accuracy of the structure that was used for the simulation. Given this, we hypothesized that we could use RMSE of the simulated to the experimental ERMS data as a proxy for predicted model accuracy (i.e., RMSD to the native structure). Thus, we rescored our generated structures according to their RMSE as described in the methods.

For each of the 20 benchmark proteins with Alphafold2-generated subunits, the RMSD of the best scoring model (of 10,000) identified using Rosetta's ref2015 score function alone was compared to the RMSD of the best scoring model after ERMS rescoring (Figure 3A). For 11 out of 20 of the proteins in our benchmark set, the best scoring model had a lower RMSD after rescoring with ERMS data (points colored blue or purple in Figure 3A). For 4/20 cases, the best scoring models improved from above 5 Å RMSD to below 5 Å RMSD (points colored blue in Figure 3A). Some models (5/20) did not significantly change (RMSDs within 0.5 Å, colored white). And finally, 4/20 were less accurate after rescoring (colored pink). However, these proteins were all inaccurate (lowest RMSD was 7.8 Å) before and after rescoring, and thus, the change was not particularly relevant. Therefore, this rescoring method drastically improved the accuracy in several (4/20, colored blue) previously poorly

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

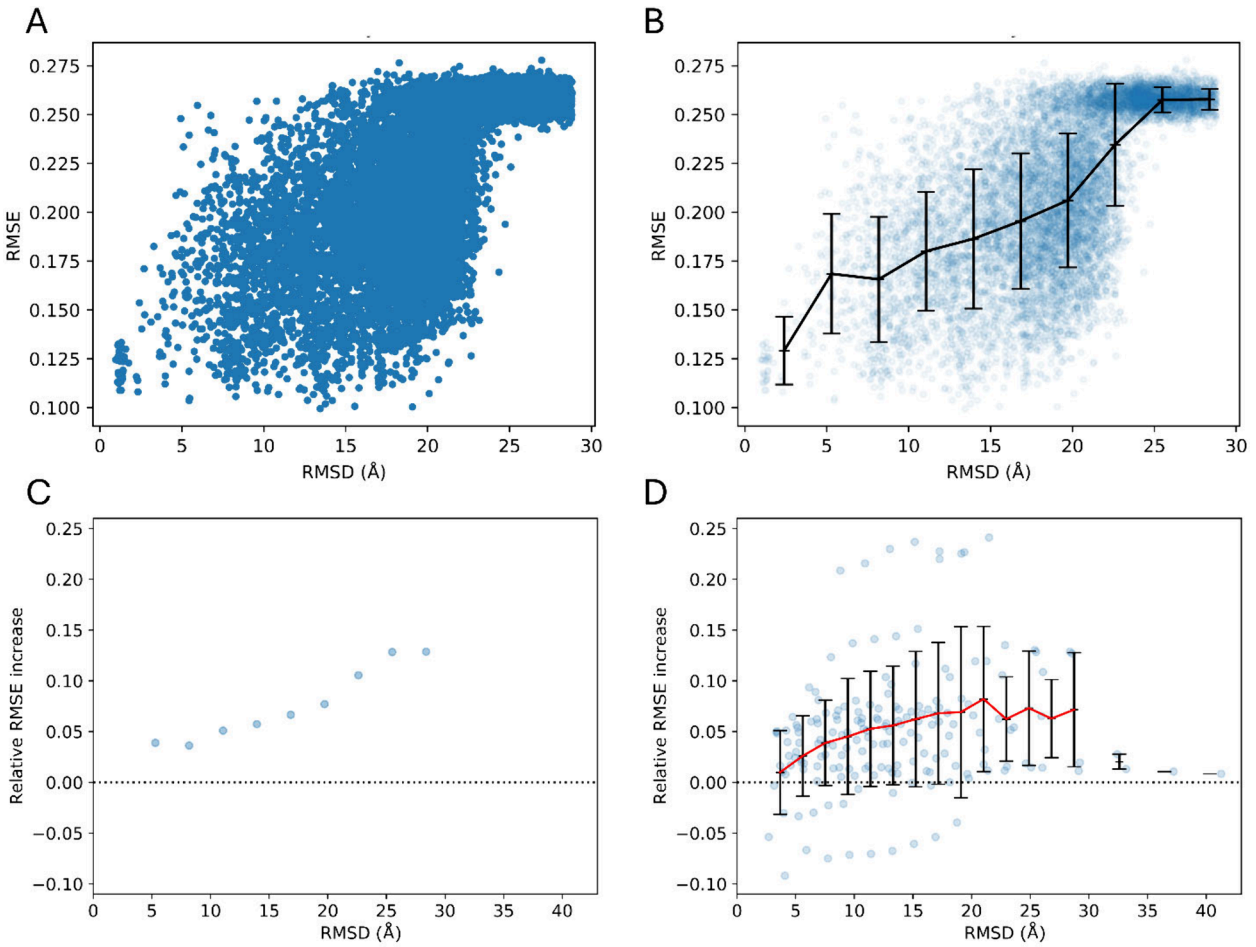


Figure 2. Correlation between accuracy of simulated SID-ERMS data ("RMSE") and accuracy of predicted protein structure ("RMSD"). Data in this figure was generated from docking of AlphaFold2-derived subunits. (A) The RMSE of the experimental ERMS to the simulated ERMS as a function of the RMSD of the predicted structure to the native structure for 10,000 RosettaDock models of the heterotetramer Tryptophan synthase (PDB: 1WBJ). (B) The RMSE/RMSD plot for 1WBJ with ten binned averages over the RMSD range. Each individual bin had RMSE averaged. (C) The relative increase in RMSE (with respect to the first bin) plotted against the RMSD for 1WBJ. (D) Relative increase in RMSE (with respect to the first bin) plotted against RMSD for all benchmark proteins. The red trendline shows the average across the entire benchmark set for bins with more than 3 points.

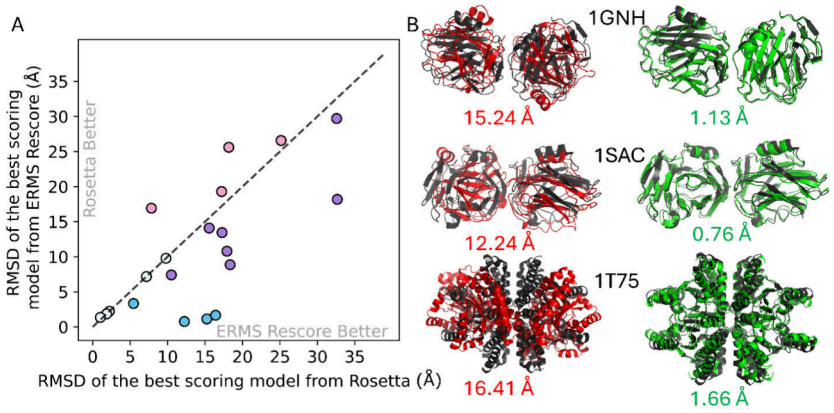


Figure 3. Comparison of Rosetta modeling and SID-ERMS modeling. (A) Comparison of the RMSD of the best scoring model using Rosetta ref2015 score and SID-ERMS rescoring for all 20 benchmark proteins. 11/20 proteins showed improvement in RMSD upon ERMS rescoring (colored blue or purple), with 4/20 exhibiting improvement below 5 Å (colored blue). Models that exhibited an RMSD change of less than 0.5 Å upon ERMS rescoring are colored white. All other models are colored pink. (B) Best scoring models of both Rosetta and SID-ERMS rescoring for three protein systems overlaid with the native structure. Red models are the Rosetta model, green models are the SID-ERMS rescore model, black is the native structure. The RMSD of each structure to the native structure is reported below the structure.

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

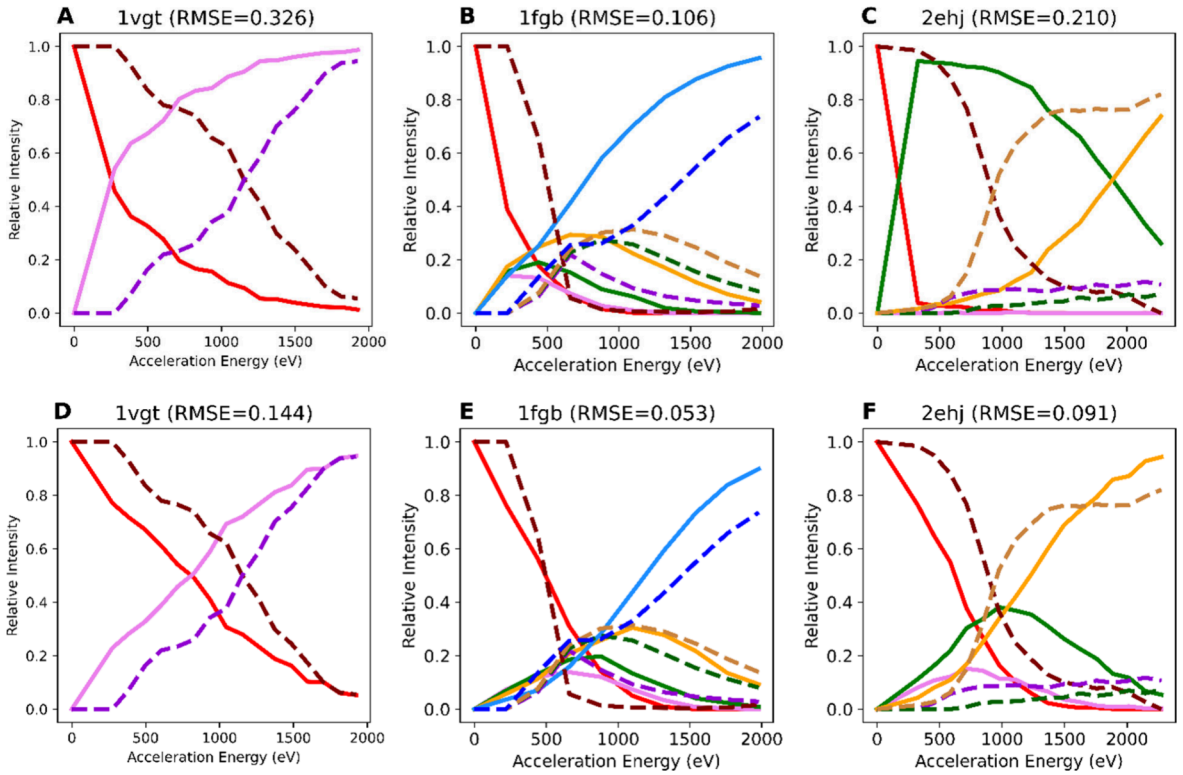


Figure 4. Comparison of the experimental ERMS data with the simulated ERMS data of the best scoring models from Rosetta modeling and SID modeling. Plots shown for proteins 1VGT (dimer), 1FGB (pentamer), and 2EHJ (tetramer). Panels A, B, C show the experimental SID-ERMS data (dashed lines) overlaid with the simulated ERMS data of the best scoring Rosetta structure. The colors of each line correspond to the oligomeric state of the complex (red = N, pink = $N-1$, green = $N-2$, yellow = $N-3$, blue = $N-4$, gray = $N-5$), where N is the stoichiometry of the unfragmented complex. Panels D, E, F show the same experimental SID-ERMS data overlaid with the simulated ERMS data of the best rescored models for 1VGT, 1FGB, and 2EHJ, respectively. The RMSE is reported for both sets of data.

predicted systems while simultaneously not worsening accurately predicted complexes (below 5 Å) in any of the test cases. To assess how sensitive our predictions were to the quality of the best scoring model, Figure S7 shows the same analysis as Figure 3, but using the average RMSD of the best three scoring models from ERMS rescoring. The distribution of points is similar to when using only the best-scoring models (Figure 3), in that 14/20 protein complexes exhibited a model matching or improving in quality upon Rosetta. However, due to the averaging, the improvement is notably worse when compared with using just the top scoring models. For example, not a single model improved to below 5 Å RMSD when the average RMSD is assessed (Figure S7). This demonstrates the robustness of our method and underscores the reason for choosing the best scoring model.

As expected, the structure of the homotetramer pyruvate kinase (PDB: 1AQF) did not improve, as the subunit prediction was inaccurate. Both before and after including ERMS data, the prediction for 1AQF was poor (over 29 Å RMSD in both cases). This further emphasizes the need for an accurate subunit structure. A similar plot showing the rescoring of complexes derived from docking PDB structure subunits can be found in Figure S8, which showed similar trends to Figure 3A. Figure 3B shows the predicted interfaces (before and after application of the ERMS data) of three proteins in the data set, two pentamers (C-reactive protein, PDB: 1GNH and Serum amyloid P, 1SAC) and a tetramer (Carbonic anhydrase, PDB: 1T75). The left structures represent the top Rosetta model overlaid on the native structure, while the right structures represent the top rescored model overlaid on the native structure. The RMSDs of these three structures are given below each structure.

Additionally, the simulated ERMS of the best scoring complexes exhibited higher agreement with the experimental ERMS data postrescoring compared to the Rosetta models (see Figure 4 for three example protein complexes). While this is somewhat intrinsic to the setup of the method, the high degree of accuracy for the best scoring models postrescoring underscored the ability for the method to produce structures in agreement with experimental ERMS data. Similar ERMS plots of the best scoring models from Rosetta and ERMS rescoring for all 20 proteins can be found in Figures S9 and S10. In summary, this demonstrated the ability of SID-ERMS data to guide the generation of accurate protein complexes, particularly in previously difficult to predict systems.

Rescoring with ERMS Data Outperformed Rescoring with SID-AE Data. In previous work, we rescored Rosetta-generated protein complexes using SID-AE (single-point) data and observed an improvement in quality of predicted models for several of our benchmark protein complexes.[13] Due to the scarcity of the SID-AE data, we hypothesized that (multipoint) ERMS data would be more useful than AE data in structure prediction. We wanted to test this hypothesis by directly comparing the results of SID-ERMS rescoring vs SID-AE rescoring. For our benchmark set, we rescored the 10,000 models using the SID-AE data instead of the SID-ERMS data, following the strategy outlined previously.[13] The RMSD of the best scoring model was recorded for the 10,000 models after rescoring with SID-AE and SID-ERMS, respectively. The results

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

of this analysis are shown in Figure 5. In 14/20 models, we identified a lower RMSD model when rescoring with ERMS data

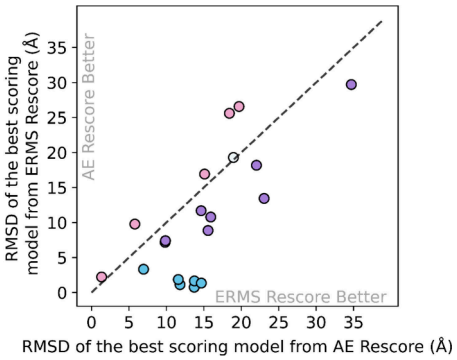


Figure 5. Comparison of SID-AE modeling and SID-ERMS modeling. Comparison of the RMSD of the best scoring model using SID-AE rescoring from previous work and the newly developed SID-ERMS rescoring. Upon using SID-ERMS data, 14/20 models showed improvement in RMSD, with 6/20 showing improvement to below 5 Å. Color pattern for the plot is identical to Figure 3.

(colored blue or purple in Figure 5). In 6/20 of these models, we saw improvement from above 5 to below 5 Å (colored blue in Figure 5). For 5/20 models, AE rescoring (colored pink) outperformed ERMS rescoring. One of those five models was highly accurate in both rescoring scenarios (RMSD = 2.22 Å for ERMS rescore and 1.35 Å for AE rescore). The other four models were all inaccurate in both rescoring scenarios (above 5 Å RMSD) and only worsened by approximately 3−7 Å, indicating that SID-AE rescoring did not significantly outperform the ERMS rescoring for that subset. These results supported our hypothesis that the ERMS data indeed constitutes a more complete source of structural information than AE. ERMS data contains a full set of relative abundance data over a range of acceleration energies, providing information on the interface strengths and breakage patterns of a protein complex at different levels of internal energy. Using the full ERMS data allows for model selection that agrees with the native complex structure across the entire acceleration energy range, resulting in higher accuracy models than using a single point of information such as AE.

CONCLUSIONS

SID-nMS is a versatile microanalytical tool for probing protein complex structure. Previous work has used single points of information from SID-ERMS data for protein complex prediction. However, no work had been done on the potential for using the complete SID-ERMS data in elucidating protein complex structure. In this work, we created a benchmark data set of 20 proteins for which we had SID-nMS ERMS data. We predicted 10,000 structures for each of the 20 proteins with RosettaDock, using both the reference structure and Alphafold2 generated subunits, respectively. We simulated the ERMS for all 200,000 structures and compared the agreement with experimental data to the accuracy of the predicted structure. We were able to show that higher RMSE generally correlated with higher RMSD, suggesting that ERMS data could be utilized to distinguish between high- and low-quality models. Hence, we developed a Rosetta score term that assessed model quality based on agreement to the experimental ERMS data. All 200,000 models were subsequently rescored using the experimental ERMS as an input, and improvement in RMSD was observed in 11/20 benchmark cases, whereas meaningful deterioration in model quality was not observed. Finally, the SID-ERMS rescored complexes were compared to the SID-AE rescored complexes, showing an improvement in RMSD in 14/20 cases. Two recurring trends observed in the nonimproving cases were beta-barrel secondary structural motifs in the subunits and overall large complex size. The beta-barrel structure could artificially lower the calculated interface strength, resulting in the predicted fragmentation occurring more readily than experimentally observed. Furthermore, it has been previously shown to be difficult to sufficiently sample docking conformations of larger complexes[13] due to the extensive conformational search space. Both of these are potential sources of improvement for future iterations of our integrative modeling method.

SID-nMS is an information rich and highly flexible method accessible to many systems. Structural and interface information can be extracted from the SID-ERMS data. Our novel SID-ERMS-guided rescoring introduces a workflow for more accurate protein complex structure elucidation in congruence with the experimental data. This method significantly outperformed the current multimeric prediction using Rosetta. It also significantly outperformed previous modeling with the SID-AE data. Our method is freely available as a Rosetta application, and we provide a tutorial with test files of the method in Supporting Information. As mentioned in the Introduction, with the lower confidence of multimeric predictions from purely computational methods like AlphaFold Multimer, SID-ERMS rescoring has the potential for higher confidence structures to be obtained. There were several proteins in our benchmark set for which we significantly improved the accuracy of the best model but still remained above 5 Å RMSD. For this reason, we speculate that SID-ERMS data in combination with machine learning complex prediction have a high potential for further improvement. Future work will focus on implementing SID-nMS data into current neural network based methods such as AlphaFold2 and RoseTTAFold 2. Additionally, secondary fragmentation has been observed at higher acceleration energies in nMS for certain complexes. Adding this functionality into the existing ERMS prediction method is another possible next step for improving structure prediction in conjunction with machine learning networks. Integration of experimental data like SID-nMS is the potential first step in bridging the gap between monomeric and multimeric structure prediction in these machine learning networks.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c05837.

Monomer RMSDs for the AlphaFold2 subunits; ERMS, RMSE, and RMSD plots for PDB structure subunits, as well as additional equivalent plots for AlphaFold2 derived subunits; tutorial for the Rosetta application (PDF)

1fgb_pdbs/: 1,000 pdb structures of 1fgb (only the dimer form is present in the pdbs for interface modeling). The models are split across 2 subdirectories; ERMS_1FGB.tsv: a.tsv file containing the experimental data for 1fgb. It uses C5 symmetry; C5_complex_type: a symmetry file describing the interfaces for C5 complexes. Used in the ERMS prediction portion of the application; 1fgb_rmsd.txt: a text file containing our RMSD

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383

calculations for all 1,000 structures; 1fgb_rmse.txt: a text file containing our RMSE calculations for all 1,000 structures; 1fgb_score.txt: a text file containing our ref2015 scorings for all 1,000 structures (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

Authors

Robert M. Bolz − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States*

Justin T. Seffernick − *Department of Structural Biology and Chemical Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States*

Zachary C. Drake − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States*

Sophie R. Harvey − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States; Native Mass Spectrometry Guided Structural Biology Center, Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0003-0763-8173

Vicki H. Wysocki − *Department of Chemistry and Biochemistry, Ohio State University, Columbus, Ohio 43210, United States; Native Mass Spectrometry Guided Structural Biology Center, Ohio State University, Columbus, Ohio 43210, United States; School of Chemistry & Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0003-0495-2538

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c05837

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank fellow Lindert lab members for their productive conversations and advice. We also thank members of the Wysocki lab for their insight into the Mass Spectrometry data. Finally, we also thank the Ohio Supercomputer Center (OSC) for their computational resources.[46] This work was supported by National Institutes of Health (RM1GM149374) and a Sloan Research Fellowship to S.L.

REFERENCES

(1) Drummond, E.; Pires, G.; MacMurray, C.; Askenazi, M.; Nayak, S.; Bourdon, M.; Safar, J.; Ueberheide, B.; Wisniewski, T. *Brain* **2020**, *143* (9), 2803−2817.

(2) Wilson, J. B.; Yamamoto, K.; Marriott, A. S.; Hussain, S.; Sung, P.; Hoatlin, M. E.; Mathew, C. G.; Takata, M.; Thompson, L. H.; Kupfer, G. M.; Jones, N. J. *Oncogene* **2008**, *27* (26), 3641−3652.

(3) Knobbe, C. B.; Revett, T. J.; Bai, Y.; Chow, V.; Jeon, A. H. W.; Böhm, C.; Ehsani, S.; Kislinger, T.; Mount, H. T.; Mak, T. W. *J. Proteome Res.* **2011**, *10* (10), 4388−4404.

(4) Shen, H.-B.; Chou, K.-C. *J. Proteome Res.* **2009**, *8* (3), 1577−1584.

(5) Kermani, A. A. *FEBS J.* **2021**, *288* (20), 5788−5804.

(6) Havel, T. F. *Prog. Biophys. Mol. Biol.* **1991**, *56* (1), 43−78.

(7) Weissenberger, G.; Henderikx, R. J. M.; Peters, P. J. *Nat. Methods* **2021**, *18* (5), 463−471.

(8) Stiving, A. Q.; VanAernum, Z. L.; Busch, F.; Harvey, S. R.; Sarni, S. H.; Wysocki, V. H. *Anal. Chem.* **2019**, *91* (1), 190−209.

(9) Leney, A. C.; Heck, A. J. R. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (1), 5−13.

(10) Quintyn, R. S.; Yan, J.; Wysocki, V. H. *Chem. Biol.* **2015**, *22* (5), 583−592.

(11) 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. U. S. A.* **2019**, *116* (17), 8143−8148.

(12) Snyder, D. T.; Harvey, S. R.; Wysocki, V. H. *Chem. Rev.* **2022**, *122* (8), 7442−7487.

(13) Seffernick, J. T.; Harvey, S. R.; Wysocki, V. H.; Lindert, S. *ACS Cent. Sci.* **2019**, *5* (8), 1330−1341.

(14) Seffernick, J. T.; Canfield, S. M.; Harvey, S. R.; Wysocki, V. H.; Lindert, S. *Anal. Chem.* **2021**, *93* (21), 7596−7605.

(15) Orbán-Németh, Z.; Beveridge, R.; Hollenstein, D. M.; Rampler, E.; Stranzl, T.; Hudecz, O.; Doblmann, J.; Schlögelhofer, P.; Mechtler, K. *Nat. Protoc.* **2018**, *13* (3), 478−494.

(16) Turzo, S. B. A.; Seffernick, J. T.; Rolland, A. D.; Donor, M. T.; Heinze, S.; Prell, J. S.; Wysocki, V. H.; Lindert, S. *Nat. Commun.* **2022**, *13* (1), 4377.

(17) Drake, Z. C.; Seffernick, J. T.; Lindert, S. *Nat. Commun.* **2022**, *13* (1), 7846.

(18) Allison, T. M.; Degiacomi, M. T.; Marklund, E. G.; Jovine, L.; Elofsson, A.; Benesch, J. L. P.; Landreh, M. *Protein Sci.* **2022**, *31* (6), No. e4333.

(19) Biehn, S. E.; Lindert, S. *Nat. Commun.* **2021**, *12* (1), 341.

(20) Britt, H. M.; Cragnolini, T.; Thalassinos, K. *Chem. Rev.* **2022**, *122* (8), 7952−7986.

(21) Kovvali, S.; Gao, Y.; Cool, A.; Lindert, S.; Wysocki, V. H.; Bell, C. E.; Gopalan, V. *Protein Sci.* **2023**, *32* (7), No. e4695.

(22) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. *Nat. Biotechnol.* **2023**, *41* (12), 1810−1819.

(23) Aprahamian, M. L.; Chea, E. E.; Jones, L. M.; Lindert, S. *Anal. Chem.* **2018**, *90* (12), 7721−7729.

(24) Bleiholder, C.; Liu, F. C. *J. Phys. Chem. B* **2019**, *123* (13), 2756−2769.

(25) Politis, A.; Park, A. Y.; Hall, Z.; Ruotolo, B. T.; Robinson, C. V. *J. Mol. Biol.* **2013**, *425* (23), 4790−4801.

(26) Degiacomi, M. T.; Schmidt, C.; Baldwin, A. J.; Benesch, J. L. P. *Structure* **2017**, *25* (11), 1751−1757.

(27) Jia, R.; Bradshaw, R. T.; Calvaresi, V.; Politis, A. *J. Am. Chem. Soc.* **2023**, *145* (14), 7768−7779.

(28) Lau, A. M.; Jia, R.; Bradshaw, R. T.; Politis, A. *Biochem. Soc. Trans.* **2020**, *48* (3), 971−979.

(29) Hauri, S.; Khakzad, H.; Happonen, L.; Teleman, J.; Malmström, J.; Malmström, L. *Nat. Commun.* **2019**, *10* (1), 192.

(30) Piotrowski, C.; Moretti, R.; Ihling, C. H.; Haedicke, A.; Liepold, T.; Lipstein, N.; Meiler, J.; Jahn, O.; Sinz, A. *Cells* **2020**, *9* (1), 136.

(31) Kahraman, A.; Herzog, F.; Leitner, A.; Rosenberger, G.; Aebersold, R.; Malmström, L. *PloS One* **2013**, *8* (9), No. e73411.

(32) Biehn, S. E.; Limpikirati, P.; Vachet, R. W.; Lindert, S. *Anal. Chem.* **2021**, *93* (23), 8188−8195.

(33) Aprahamian, M. L.; Lindert, S. *J. Chem. Theory Comput.* **2019**, *15* (5), 3410−3424.

(34) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. *Anal. Chem.* **2010**, *82* (22), 9557−9565.

(35) Mitra, R.; Usher, E. T.; Dedeoğlu, S.; Crotteau, M. J.; Fraser, O. A.; Yennawar, N. H.; Gadkari, V. V.; Ruotolo, B. T.; Holehouse, A. S.; Salmon, L. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121*, e2409139121.

(36) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268* (1), 209−225.

(37) Seffernick, J. T.; Turzo, S. B. A.; Harvey, S. R.; Kim, Y.; Somogyi, Á.; Marciano, S.; Wysocki, V. H.; Lindert, S. *Anal. Chem.* **2022**, *94* (29), 10506−10514.

(38) Stranges, P. B.; Kuhlman, B. *Protein Sci.* **2013**, *22* (1), 74−82.

(39) 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. *Nature* **2021**, *596* (7873), 583−589.
(40) 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.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1711−1721.
(41) Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. *Bioinformatics* **2023**, *39* (7), btad424.
(42) Arslanian, A. J.; Wysocki, V. H. *J. Am. Soc. Mass Spectrom.* **2025**, *36* (1), 187−200.
(43) Sarni, S. H.; Roca, J.; Du, C.; Jia, M.; Li, H.; Damjanovic, A.; Małecka, E. M.; Wysocki, V. H.; Woodson, S. A. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (47), No. e2208780119.
(44) 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. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031−3048.
(45) *The PyMOL Molecular Graphics System*, Version 1.8; ,Schrödinger, LLC, 2015.
(46) Ohio Supercomputer Center, 1987.

https://doi.org/10.1021/acs.analchem.4c05837
Anal. Chem. 2025, 97, 2375−2383