Article

Prediction of Protein Complex Structure Using Surface-Induced Dissociation and Cryo-Electron Microscopy

Justin T. Seffernick, Shane M. Canfield, Sophie R. Harvey, Vicki H. Wysocki, and Steffen Lindert*

Read Online

| ACCESS | | Metrics & More | | Article Recommendations | | Supporting Information |

ABSTRACT: A variety of techniques involving the use of mass spectrometry (MS) have been developed to obtain structural information on proteins and protein complexes. One example of these techniques, surface-induced dissociation (SID), has been used to study the oligomeric state and connectivity of protein complexes. Recently, we demonstrated that appearance energies (AE) could be extracted from SID experiments and that they correlate with structural features of specific protein–protein interfaces. While SID AE provides some structural information, the AE data alone are not sufficient to determine the structures of the complexes. For this reason, we sought to supplement the data with computational modeling, through protein–protein docking. In a previous study, we demonstrated that the scoring of structures generated from protein–protein docking could be improved with the inclusion of SID data; however, this work relied on knowledge of the correct tertiary structure and only built full complexes for a few cases. Here, we performed docking using input structures that require less prior knowledge, using homology models, unbound crystal structures, and bound+perturbed crystal structures. Using flexible ensemble docking (to build primarily subcomplexes from an ensemble of backbone structures), the $RMSD_{100}$ of all (15/15) predicted structures using the combined Rosetta, cryo-electron microscopy (cryo-EM), and SID score was less than 4 Å, compared to only 7/15 without SID and cryo-EM. Symmetric docking (which used symmetry to build full complexes) resulted in predicted structures with $RMSD_{100}$ less than 4 Å for 14/15 cases with experimental data, compared to only 5/15 without SID and cryo-EM. Finally, we also developed a confidence metric for which all (26/26) proteins flagged as high confidence were accurately predicted.
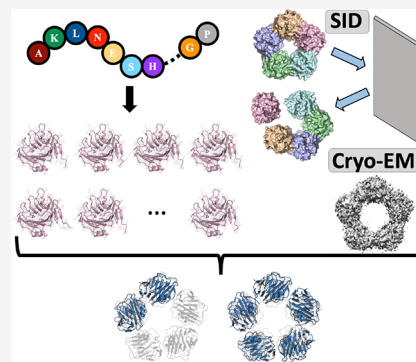
INTRODUCTION

Mass spectrometry (MS) can be used to elucidate elements of protein structure using a variety of techniques. This is of great significance because the determination of protein structure can markedly facilitate the development of new therapeutics through a variety of different approaches.[1] Furthermore, structural knowledge of protein complexes is particularly important because approximately 86% of proteins interact with other proteins to form complexes in vivo.[2] MS provides many advantages such as the ability to collect data using small sample sizes (typically μLs of the sample at low μM concentrations), on complex mixtures, and on both small and large protein systems (up to megadalton).[3–5] Some examples of MS-based methods that can provide structural information for proteins and complexes are surface-induced dissociation (SID),[6–13] chemical cross-linking (XL),[14–18] covalent labeling (CL),[19–24] and ion mobility (IM).[25–28] However, the data collected from MS experiments are insufficient to fully elucidate protein structure, although the information gained (masses of the intact complex and its subunits, post-translational modification [PTM] information, stoichiometry and topology, ligands bound) can be informative and can help guide or allow better interpretation of data from other structural biology tools.

High-resolution experimental protein structure determination approaches such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) exist; however, each method has significant shortcomings with current technology. With the exception of cryo-EM, these methods are especially challenging when considering large complexes, and thus alternative approaches such as native MS must frequently be used. As alternatives, some of the higher resolution approaches can be used at lower resolutions, such as when cheaper, less extensive NMR experiments are performed,[29–31] when low-resolution density maps are obtained by cryo-EM,[32–36] or when a technique known to be lower resolution is used, such as small-angle X-ray scattering (SAXS),[37–39] etc. However, the sparse data obtained from these types of experiments are typically incomplete, noisy, or sometimes inaccurate.
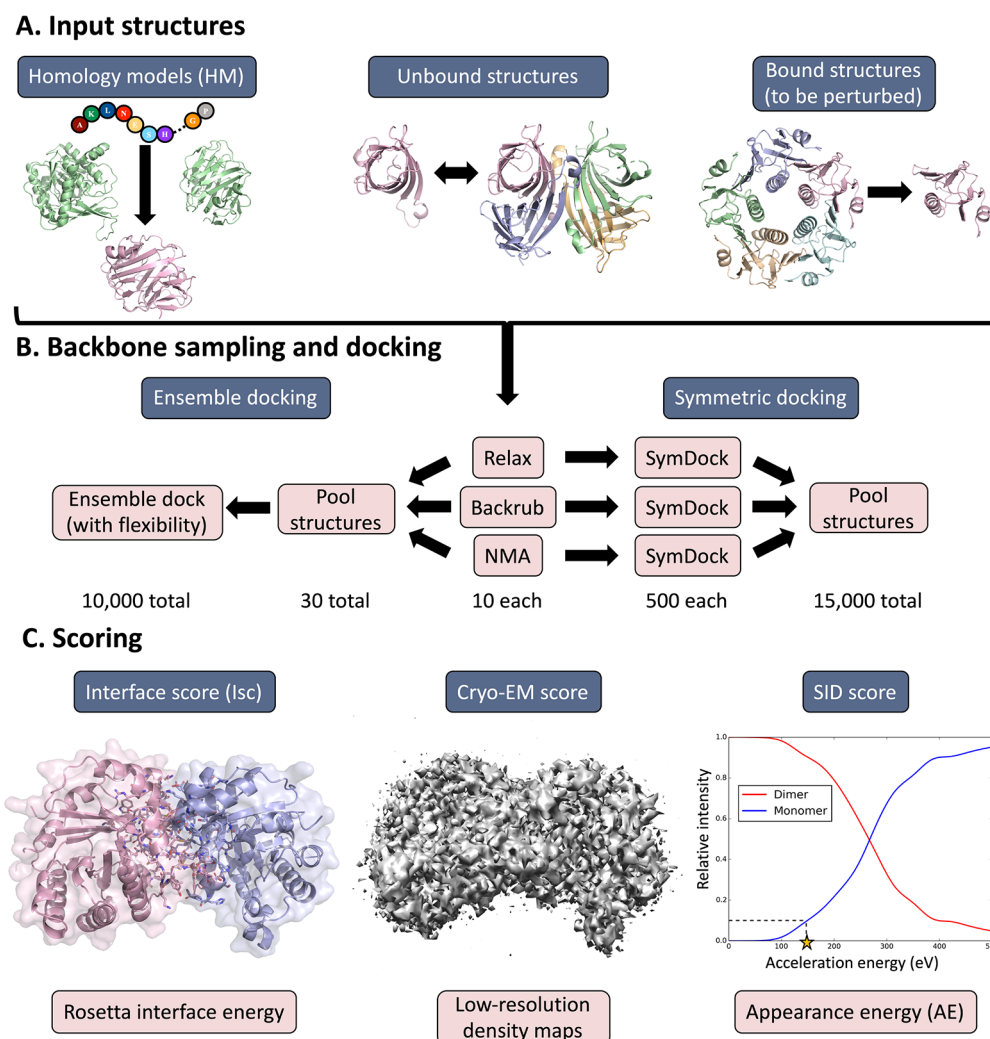
Figure 1. Overview of docking approach. (A) Input structures for docking were acquired by building homology models, obtaining unbound crystal structures, and extracting structures from bound crystal structures (followed by backbone perturbation in panel B). (B) Extra backbone sampling was performed on all input structures via Rosetta relax, backrub, and normal mode analysis (NMA), respectively. Ensemble docking (left) was performed by pooling the different backbones and docking with flexibility. Symmetric docking (right) was performed rigidly but with each backbone structure as a starting point for separate trajectories. (C) Scoring was performed using three terms: Rosetta interface score (Isc, score without including experimental data), Rosetta cryo-EM score using noisy, low-resolution density maps, and the appearance energy (AE)-dependent SID score.

While data obtained from these lower resolution methods are structure-dependent, the provided information is not enough to fully determine the coordinates of all atoms in the protein complex. One way to extract more structural information from these data is to combine them with computational methods. Many different approaches that use computational methods to supplement the information obtained from sparse data for structure prediction have been developed.[13,16−18,20−24,26−44] Computational methods to predict tertiary structure from sequence have been developed as have protein−protein docking methods, which can be used to predict quaternary structure specifically when working with protein complexes.

As previously stated, MS can be used to obtain structural information on proteins. While many methods have been developed to predict structures using XL, CL, and IM, we have demonstrated that surface-induced dissociation (SID) can be used to collect structural data as well, specifically for protein

complexes.[11,12] In SID, protein complexes are softly ionized in the gas phase and accelerated toward a surface. Upon collision with the surface, the complexes can dissociate into monomers or subcomplexes, breaking the noncovalent interactions of the interfaces, and the resulting products are analyzed with MS. The experiment can be repeated over a range of acceleration energies to obtain an overall pattern of breakage, resulting in energy-resolved mass spectrometry (ERMS) plots. From these analyses, SID has been used to extract information on stoichiometry and connectivity.[10,45,46] While SID has been incorporated into multiple different instrument platforms from different vendors in work that is ongoing, including quadrupole time-of-flight, Orbitrap, and FT ICR instruments, this has been limited previously to in-house modified instruments plus a few beta test situations.[47−50] Recently, SID has become commercially available on a high-resolution ion mobility-quadrupole time-of-flight instrument (Waters Select Series Cyclic IMS)

increasing the potential applications of SID for protein structural studies by a broader range of users.

We have previously demonstrated that weaker interfaces break more frequently than stronger interfaces at low acceleration energies.[12] By extracting the appearance energy (AE, lab-frame energy at which specific interfaces begin to break, defined in our previous work as based on 10% intensity of the precursor), a link between SID data and protein structure was determined. (This arbitrary value of 10% $[E_{10}]$ was used to avoid odd onset behavior associated with nonuniform distributions of energy in different precursors prior to surface collision and different onset slopes. We wanted to avoid use of E_{50} for our particular application because at 50% dissociation, multiple competitive pathways can overlap for higher oligomeric states of a given complex.) We hypothesized that interfaces with more favorable interactions (i.e., large interface area, more hydrogen bonds, salt bridges, etc.) would result in higher AE. From this, we developed models to predict the AE from structure. Using an AE prediction method, we showed that the information contained in the model was sufficient to discriminate between nativelike and non-nativelike structures.[13] To do this, we redocked bound crystal structures to generate a set of complex structures using RosettaDock, which varied significantly in structural accuracy. Using a scoring function derived from the difference between predicted and experimental AE for each structure, we showed that the SID AE data could indeed facilitate the discrimination between good and bad poses. While these previous results were very encouraging, ultimately, there were a few remaining shortcomings. Examples of shortcomings are as follows: we used tertiary structures from crystal structures of the complexes as input into the docking, we performed rigid docking only, we only attempted to build structures of entire complexes in 3/9 cases, we identified success based on obtaining a good structure in the top 3 scoring models, and finally, accurate structures were only identified in ~66% of cases.

In addition to SID, cryo-EM can be used to obtain sparse data for protein complexes. While it is possible to obtain high-resolution density maps from cryo-EM, sometimes only low-resolution density maps are reconstructed, requiring further supplementation using computational methods. We hypothesized that data obtained from SID and low-resolution cryo-EM would be complementary for a variety of reasons. For one, both methods are much higher throughput than full structure determination using either X-ray crystallography or NMR as they require relatively small amounts of sample. Additionally, both methods are compatible with large protein complexes, while X-ray crystallography and NMR have significant size limitations, making them incompatible for very large systems. Finally, we hypothesized that SID and cryo-EM provide orthogonal structural information. SID AE is dependent on interface composition, while cryo-EM low-resolution density maps are dependent on the overall shape of the complex.

In this study, we performed a plethora of docking simulations to demonstrate the ability of SID to accurately predict full complex structures in realistic scenarios, essentially overcoming the noted shortcomings from our previous docking study. Here, we combined SID AE with simulated cryo-EM low-resolution density maps to predict structures of protein complexes using protein–protein docking. Rather than simply using bound crystal structures, i.e., structures in which all monomers are bound to form a complex, which required

knowledge of the complex structure for input into docking, we obtained input tertiary structures in multiple different ways: homology models (HM), unbound (monomer or subcomplex) crystal structures, and bound+perturbed crystal structures. We then performed docking on these input structures two different ways. First, we performed ensemble docking using Rosetta-Dock, which allowed for a flexible backbone during the docking based on an ensemble of generated input structures. Next, we performed rigid symmetric docking using Rosetta SymDock[51] with the same ensemble of input structures to build whole complexes. This symmetric docking approach was the first docking strategy with SID to routinely build full complexes. Scoring and model selection were based on scores from Rosetta, cryo-EM, and SID. An overview of the method is shown in Figure 1. For ensemble docking, the $RMSD_{100}$ of all (15/15) predicted structures using the combined Rosetta, cryo-EM, and SID score was less than 4 Å, compared to 7/15 without including experimental data. For symmetric docking, the $RMSD_{100}$ of the predicted, full complex structures using the combined score was less than 4 Å for 14/15 cases, compared to 5/15 without cryo-EM and SID.

■ METHODS

Data Set. The data set used for protein–protein docking with SID was described previously.[13] In short, the data set contained triose phosphate isomerase (homodimer, 8TIM), streptavidin (homotetramer, 1SWB), hemoglobin (heterotetramer, 1GZX), cholera toxin B (homopentamer, 1FGB), C-reactive protein (homopentamer, 1GNH), and serum amyloid P (homopentamer, 1SAC). For the complexes in the data set, experimental SID AE values for specific interfaces were extracted from ERMS plots and normalized by the number of intersubunit protein–protein interfaces as described previously.[12,13] Simulation of the 14 Å resolution density maps with noise is described in the SI.

Input Structures for Docking. Rather than solely redocking crystal structures for full complexes, here, we obtained docking input structures three different ways, which are visually depicted in Figure 1A and described in further detail in the SI. In short, when possible, we built homology models, used unbound crystal structures, and extracted bound crystal structures.

Rather than performing completely rigid docking simulations with single input structures, we performed extra backbone sampling for all inputs prior to docking. For each input structure, 10 models were built using Rosetta relax, normal mode analysis (NMA), and backrub, resulting in a total of 30 models with slightly different backbones. These structures were input into docking as described in the following sections.

Ensemble Docking. Using a somewhat similar approach to our previous rigid docking method with SID,[13] we performed docking in order to build (sub)complexes corresponding to each interface where AE was extracted. However, here we performed ensemble docking, where backbones were allowed to swap during the simulation from a pool of inputs. As described in the SI, these docking simulations were performed for all of the different types of input structures: 5 for HM, 2 for unbound, and 8 for bound+perturbed. A list of the specific interfaces and partners built for each input structure type is provided in Table S1. Additionally, for each input structure, the 30 extra backbone structures were pooled prior to docking. The orientation of the

mobile chain was randomized using the -randomize2 flag. Ensemble docking in RosettaDock[52] was performed where the prepacked monomers were allowed to swap during the simulation, allowing for backbone flexibility based on the sampled tertiary structures. For each input structure, 10,000 docked structures were built. The workflow for ensemble docking is shown in Figure 1B (left).

Symmetric Docking. In addition to ensemble docking, symmetric docking was also performed to build full complexes in all cases based on the symmetry derived from stoichiometry provided by SID data. De novo symmetry was used when possible; however, due to poor observed sampling when using D2 symmetry, complexes for streptavidin (1SWB and 5N8T) were docked using noncrystallographic point symmetry. A list of applied symmetries for each input structure is provided in Table S2. Note that for hemoglobin (1GZX, dimer of heterodimers) in both HM and bound+perturbed, the input heterodimer was the predicted structure from the ensemble docking (using the combined score described in the following section) and was subsequently docked with C2 symmetry to form the heterotetramer. A definition of C2 symmetry was also used in all other cases where dimers were used to build tetramers (2HBC, 1GZX_dimers, and 1SWB_dimers). For symmetric docking, backbones of input structures were not allowed to swap during the simulations; therefore, separate trajectories of 500 structures for each of the 30 inputs were performed for a total of 15,000 docked structures. The workflow for symmetric docking is shown in Figure 1B (right).

Scoring Strategy. Structures from both the ensemble and symmetric docking were scored using the Rosetta interface score (Isc). Then, cryo-EM and SID scores were subsequently included in the overall scoring. The components corresponding to the score terms are visually depicted in Figure 1C. Rosetta Isc represents the Rosetta energy of the interface, and the model with the lowest Isc represents the predicted model using Rosetta without experimental data. The Rosetta cryo-EM score (elec_dens_fast) was calculated based on the generated low-resolution density maps with noise.[36,53,54] Calculation of SID scores was described in-depth previously.[13] In short, for each docked structure, the AE was predicted using the model (with rigidity factor [RF][12] specifically calculated and averaged for the ensemble of input structures) and compared to the experimental AE. Structures were scored based on this difference using a fade function, with no penalty assigned to structures for which the predicted AE was within 100 eV of the experimental lab frame energy. If multiple AEs were available for a complex where all interfaces corresponding to the aforementioned AEs were built, the SID score used was the average SID score for those AEs. For the combined scores (Isc +cryo-EM and Isc+cryo-EM+SID), each individual score was first normalized based on the maximum observed absolute value for the score of the docked structures for that protein. The total score was a weighted sum of the three normalized terms, as shown in eq 1. The weights (w_{Isc}, $w_{cryo-EM}$, and w_{SID}) used for ensemble and symmetric docking are provided in Table S3. We always chose the lowest scoring model as the predicted structure.

$$Score = w_{Isc}Isc + w_{cryo-EM}Score_{cryo-EM} + w_{SID}Score_{SID} \quad (1)$$

Confidence Metric. A confidence metric was calculated in order to classify the prediction for each protein as either high or low confidence without the knowledge of the native structure. To derive the confidence metric, we first calculated RMSD (all atom, refined, and aligned) of each model with respect to the top overall scoring model for each protein. The confidence metric was defined as the P_{near} ($\lambda = 10.0$ and $k_BT = 2.0$, using the RMSD values with respect to the top scoring model). We hypothesized that when accurate predictions ($RMSD_{100}$ of top scoring model less than 4 Å) were made, a larger number of favorably scoring models would be more similar to the top scoring model and superior funneling would be observed, thus P_{near} (with respect to the top scoring model) would be high. Proteins for which P_{near} was greater than 0.1 were identified as high confidence.

■ RESULTS AND DISCUSSION

Previous Work Showed SID AE Could Improve Model Selection. While SID has typically been used to determine stoichiometry and connectivity for protein complexes, we established previously that a structurally dependent measure, appearance energy (AE), for protein−protein interfaces could be extracted from SID.[12] We observed correlations between AE and interface features such as size and other measures relating to interface energy in Rosetta. From these correlations, methods to predict AE from the structure were developed. AE prediction methods were then used to score sets of docked structures based on the agreement between predicted and observed AE for each structure.[13] We showed that the structural information contained in the prediction model was sufficient to facilitate the discrimination between nativelike and non-nativelike poses. When the SID scoring function was used, for 3/9 cases, the RMSD of the selected model improved by more than 18 Å with respect to when no experimental data were included, and the inclusion of SID was never detrimental. We observed that the SID scoring function generally scored a variety of structures equally well (notably including many of the top models) but penalized the majority of bad structures. While this study laid the foundation of SID usage in protein complex structure prediction, showing for the first time that SID data could facilitate model selection for a diverse set of docked poses, there were a few shortcomings in our initial implementation that we addressed here. Originally, the tertiary structure obtained from crystal structures of the complexes was used as input into the rigid docking. Additionally, for the majority of our docking simulations (6/9), we predicted structures of subcomplexes rather than entire complexes. For these reasons, we sought to perform a set of docking simulations that was much more realistic, by modifying the input structures to decrease the amount of knowledge required and also building structures of the entire complexes in more cases.

Different Strategies for Obtaining Input Structures for Docking Require Less Prior Knowledge. To perform more realistic docking simulations and thus reduce the necessary prior knowledge for docking by not relying on the bound structures, we performed docking using alternative input structures. First, when possible, we built homology models of monomers. The monomer building blocks were accurately built (RMSD < 2 Å) using RosettaCM multi-template modeling for the following five proteins: 1GNH (RMSD of 1.36 Å), 1GZX (chain A: 1.70 Å, chain B: 1.74 Å), 1SAC (1.10 Å), 1SWB (0.59 Å), and 8TIM (1.12 Å). Next, we used unbound structures for the following two proteins: hemoglobin (heterodimer, 2HBC, corresponding to a sub-structure of 1GZX) and streptavidin (monomer, 5N8T, corresponding to a subunit of 1SWB). Unbound input
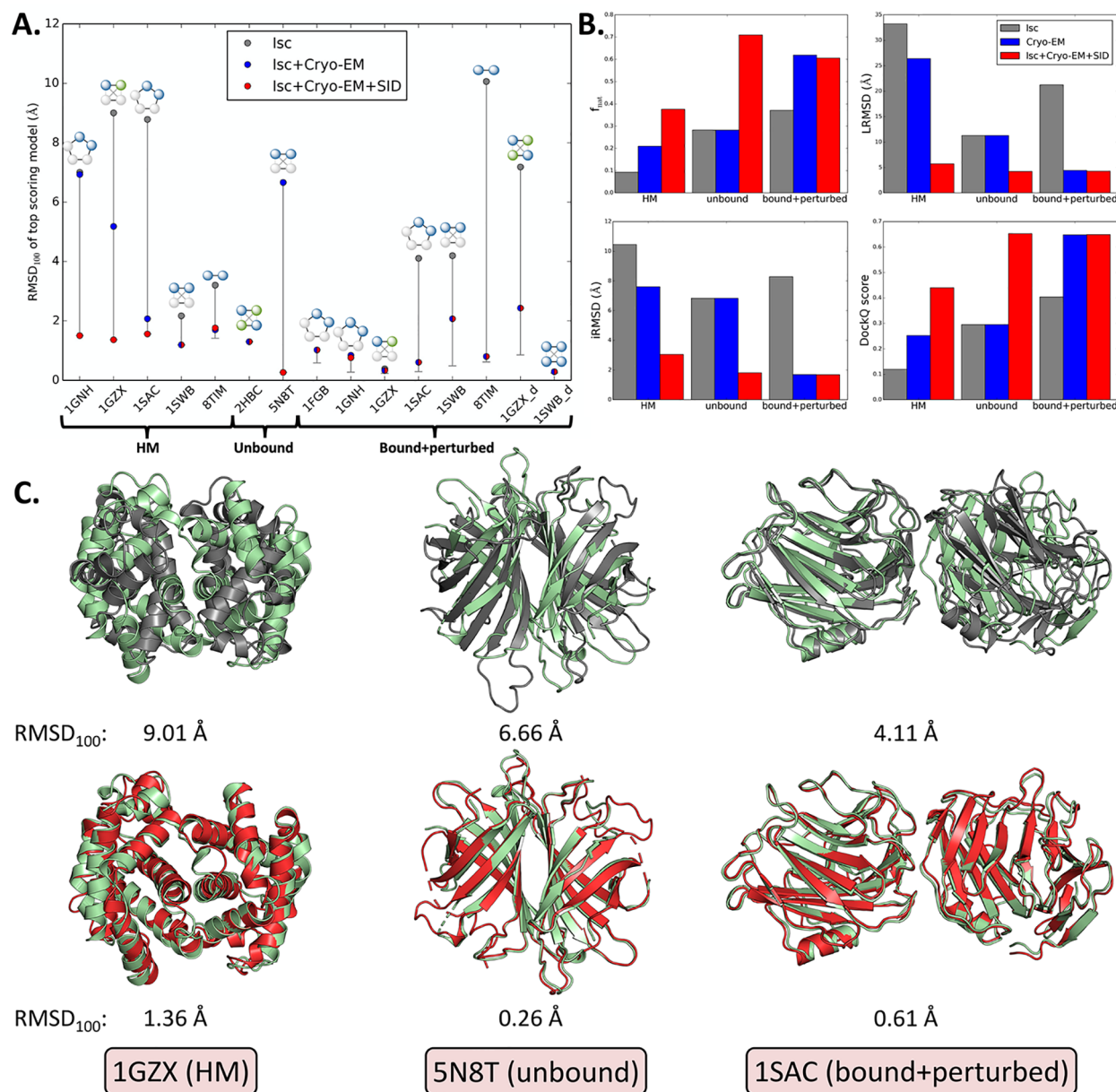
Figure 2. (A) All atom, aligned $RMSD_{100}$ of the top scoring models for ensemble docking when including no experimental data (Isc, gray), using cryo-EM data (blue), and when including both SID and cryo-EM (red). Cartoon complexes are shown for each, with the portions built in each simulation depicted in color. The bottom of the solid gray lines indicates the overall best model built. Gray points are always on top of the vertical gray lines. In all cases, $RMSD_{100}$ of the predicted structure using the combined score was less than 4 Å (only 7/15 without experimental data). (B) Dependence on input structure category. Metrics shown are f_{nat}, LRMSD, iRMSD, and DockQ score. Overall, the results were best with all experimental data. In general, bound+perturbed performed the best followed by unbound and HM, respectively. (C) Predicted structures for three cases (1GZX, HM; 5N8T, unbound; 1SAC, bound+perturbed) without (gray) and with (red) experimental data (native shown in green).

structures were not available for the remaining proteins in the benchmark set. Unsurprisingly, the RMSD of both unbound input structures (with respect to bound) was small (1.48 Å for 2HBC to 1GZX and 0.51 Å for 5N8T to 1SWB). Finally, for all proteins in the data set, we perturbed the backbones of the bound structures using several backbone-altering methods so as to not run the simulation with the tertiary structure that exactly matched the bound crystal structures. In addition to adjusting the backbones of bound structures, the same approach was used for all HM and unbound input structures with the purpose of adding flexibility into all docking as well. For each input structure, different backbones were sampled by

generating 10 models each using Rosetta relax, backrub, and normal mode analysis (NMA), for a total of 30 models. Overall, when using these sampling methods, the structures did not change drastically, with RMSDs comparing inputs and outputs of typically ∼0.5 Å and rarely above 1.5 Å, as shown in Figure S1. These ensembles of structures with slightly different backbone conformations were input into docking as described in the following sections.

Scoring Strategy Combines SID and Cryo-EM. In addition to using data from SID to score structures, here, simulated low-resolution cryo-EM density maps with noise were also included in scoring. To evaluate the complex
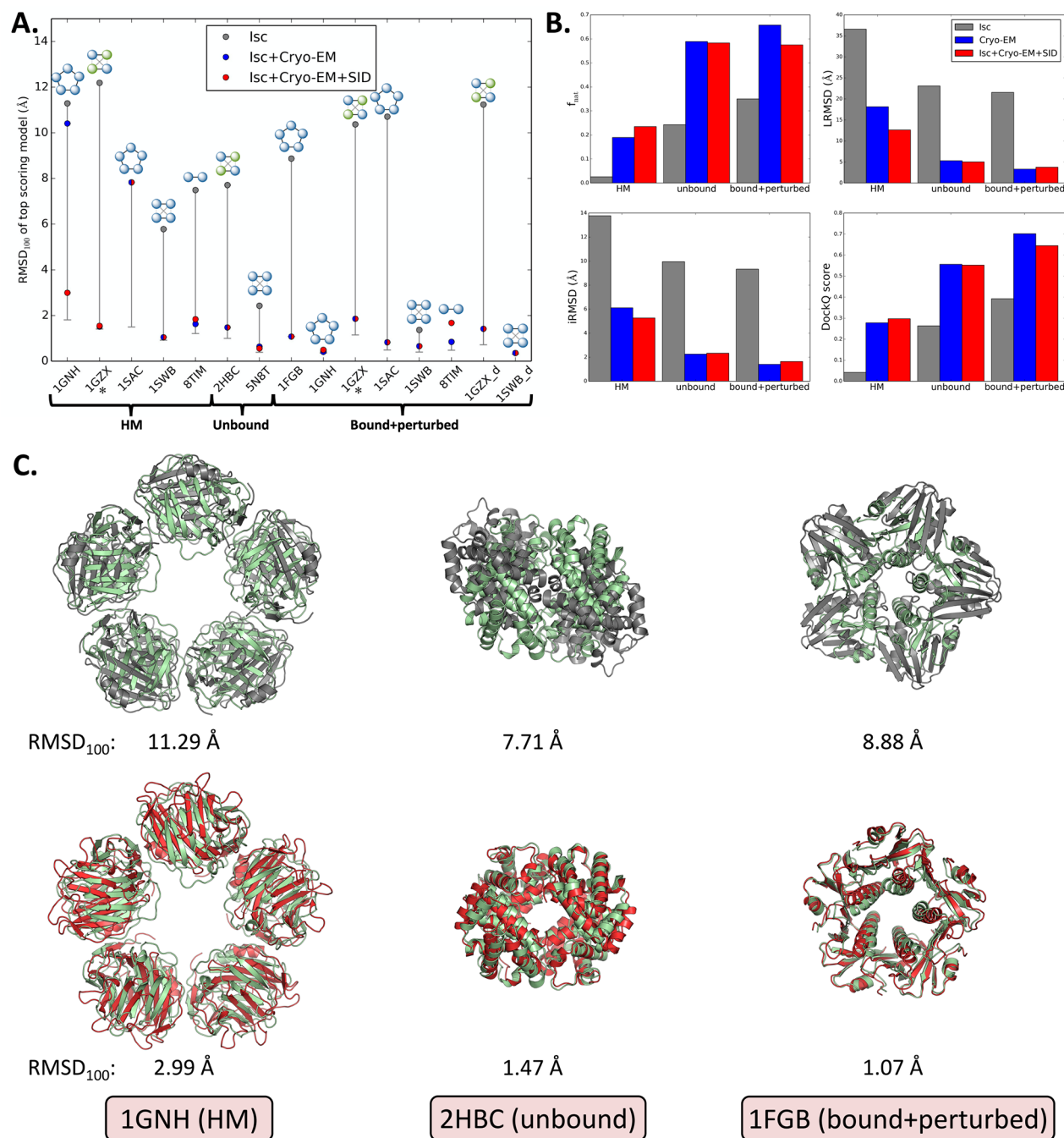
Figure 3. (A) All atom, aligned $RMSD_{100}$ of the top scoring models for symmetric docking when including no experimental data (Isc, gray), using cryo-EM data (blue), and when including both SID and cryo-EM (red). Input structures were obtained from ensemble docking for two cases (indicated with an asterisk). Cartoon complexes are shown for each. All symmetric docking simulations resulted in full complexes. The bottom of the solid gray lines indicates the overall best model built. Gray points are always on top of the vertical gray lines. In 14/15 cases, $RMSD_{100}$ of the predicted structure using the combined score was less than 4 Å (only 5/15 without experimental data). (B) Dependence on input structure category. Metrics shown are f_{nat}, LRMSD, iRMSD, and DockQ score. Typically, the average results were best with all experimental data. In general, bound+perturbed performed the best followed by unbound and HM, respectively. (C) Predicted structures for three cases (1GNH, HM; 2HBC, unbound; 1FGB, bound+perturbed) without (gray) and with (red) experimental data (native shown in green).

structure prediction when including experimental data, scoring was performed using three different scoring functions: Rosetta interface score (Isc, i.e., score with no experimental data included), Isc with Rosetta cryo-EM score, and a combined Isc, cryo-EM, and SID score. The combined score was a linear combination of the three normalized terms and represented the score when cryo-EM and SID data were both included.

The weights used for Isc, cryo-EM, and SID were the same for each protein; however, different sets of weights were used for ensemble and symmetric docking (HM and crystal). The predicted structure for each scoring function was identified as the top scoring model. The ensemble and symmetric docking are described in detail in the following sections, but we also compared the score vs RMSD plots for the SID scores

individually to previous work. Example SID score vs RMSD plots for 1SWB are shown in Figure S2 for bound rigid redocking (previous work),[13] HM ensemble docking, unbound (5N8T) ensemble docking, bound+perturbed ensemble docking, HM symmetric docking, unbound (5N8T) symmetric docking, and bound+perturbed symmetric docking. We note that while SID scores were not normalized in previous work and RMSD values were calculated differently (they were essentially ligand RMSD [LRMSD] values), the results are generally comparable. In all cases, the general trend that most accurate (low RMSD) structures were scored favorably and most inaccurate structures (high RMSD) were scored unfavorably continued to be observed. Although the SID scores are not enough to predict structures, they can be beneficial when combined with other scores (such as Isc and cryo-EM scores described here), when structures are scored slightly incorrectly. This example demonstrates that the relationship between SID AE and interface composition was similar here to previous work.

Ensemble Docking Resulted in Accurate Modeling of Subcomplexes, Strengthened by SID Data. To increase the conformational space sampled during docking, we performed flexible ensemble docking (as opposed to rigid docking) using RosettaDock. This docking was performed for all input structures (5 for HM, 2 for unbound, and 8 for bound+perturbed; see Table S1) and was used to build complexes across all interfaces with measured experimental AE. Figure 2A shows the $RMSD_{100}$ of the top scoring model using each scoring function. When no experimental data were included (Isc, gray), the $RMSD_{100}$ of the predicted model was less than 4 Å for only a few cases (2/5 for HM, 1/2 for unbound, and 4/8 for bound+perturbed). When cryo-EM data were included in scoring (Isc+cryo-EM), the $RMSD_{100}$ of the predicted structure was less than 4 Å for 12/15 cases. After the inclusion of SID data (Isc+cryo-EM+SID), all (15/15) cases resulted in accurate prediction. For predictions with this level of accuracy, the method was able to identify the correct complex topology. Remarkably, for five cases, the lowest RMSD model was also the top scoring model for the combined Rosetta, cryo-EM, and SID score. Predicted structures (and $RMSD_{100}$) are shown in Figure 2C (aligned to native shown in green) for 1GZX (HM), 5N8T (unbound), and 1SAC (bound+perturbed). Agreement with the native crystal structure improved significantly when all experimental data were included.

The average model quality of the predicted structure generally improved with the inclusion of additional experimental data (both cryo-EM and then SID) when considering the CAPRI (Critical Assessment of PRediction of Interactions) metrics, as shown in Table S4 (left) and Figure 2B. For example, the average DockQ score,[55] a quantitative measure of model quality inspired by CAPRI categories, improved from 0.30 when no experimental data were included to 0.48 when cryo-EM was included and further improved to 0.58 when SID data were also included. The same trend was observed for average RMSD (all atom, aligned), $RMSD_{100}$, f_{nat}, $f_{non-nat}$, iRMSD, and LRMSD. Using CAPRI model quality designations, 9/15 predicted structures using the combined score were identified as medium or high quality, with no cases identified as incorrect (i.e., all were at least acceptable quality), compared to 9/15 incorrect when scoring without experimental data. Improvement of funneling for the score vs RMSD distributions was also observed when experimental data were included, with the P_{near} improving by an average of 0.27 (on a

scale of 0 to 1). The RMSD for the top 50 scoring models improved by an average of 2.8 Å when SID and cryo-EM data were included as well. While much of the improvement when experimental data were included was due to the inclusion of the cryo-EM score, the SID score did play a significant role in a few cases, namely 1GNH (HM), 1GZX (HM), and 5N8T (unbound). Furthermore, accurate structures were predicted in all cases.

Symmetric Docking Resulted in Accurate Modeling Overall. In order to predict structures of entire complexes, rather than building only across interfaces identified with AE, we performed symmetric docking using Rosetta SymDock. In comparison to our previous docking studies, the symmetric docking was new and extensively increased the breadth of SID-guided modeling. In addition to each input structure (5 for HM, 2 for unbound, and 8 for perturbed bound; see Table S2), as the dimer of heterodimers could not be directly built with the structures of the monomers and symmetric docking, input structures for two cases were obtained from the output of the ensemble docking (1GZX HM and bound+perturbed, with $RMSD_{100}$s of 1.36 and 0.33 Å, respectively). Figure 3A shows $RMSD_{100}$ of the top scoring model using each scoring function. When no experimental data were included (Isc, gray), the $RMSD_{100}$ of the predicted model was less than 4 Å for only 5/15 cases (0/5 for HM, 1/2 for unbound, and 4/8 for bound+perturbed). When data from both SID and cryo-EM were included, the $RMSD_{100}$ of the predicted structure was less than 4 Å for 14/15 cases (4/5 for HM, 2/2 for unbound, and 8/8 for perturbed bound). Predicted structures are shown in Figure 3C (aligned to native shown in green) for 1GNH (HM), 2HBC (unbound), and 1FGB (bound+perturbed). Agreement with the native crystal structure improved significantly when all experimental data were included for these cases.

When considering the CAPRI metrics, as shown in Table S4 (right) and Figure 3B, the results were generally best when cryo-EM and SID data were included. The averages for all RMSD-based metrics improved when more data were included. The average results appeared slightly worse for f_{nat}, $f_{non-nat}$, and DockQ score upon the inclusion of SID (comparing Isc+cryo-EM to Isc+cryo-EM+SID); however, this was due to a single case, 8TIM (average DockQ score improves slightly from 0.54 to 0.55 for the other proteins). To illustrate this minor, nondeleterious difference in the structure for 8TIM, the structures predicted from Isc+cryo-EM and Isc+cryo-EM+SID are shown in Figure S3. Noticeably, there are very few differences in the accuracy of the predicted structure. When assessing model quality using DockQ score, 10/15 models were identified as inaccurate without experimental data. However, when the combined score including SID was used, 7/15 were identified as medium or high quality (only two inaccurate). For symmetric docking, improvement in funneling was observed when SID and cryo-EM data were included. P_{near}, goodness of funneling in score vs RMSD distributions, improved over Isc by an average of 0.37 (on a scale of 0 to 1). Improvement of the average RMSD of 7.3 Å for the top 50 scoring models was also observed when SID and cryo-EM data were included. In summary, the symmetric docking results were slightly improved by SID, but the inclusion of both cryo-EM and SID produced the best results. Importantly, for the first time, all full complexes were built using SID data, and input structures required much less knowledge than previous work.
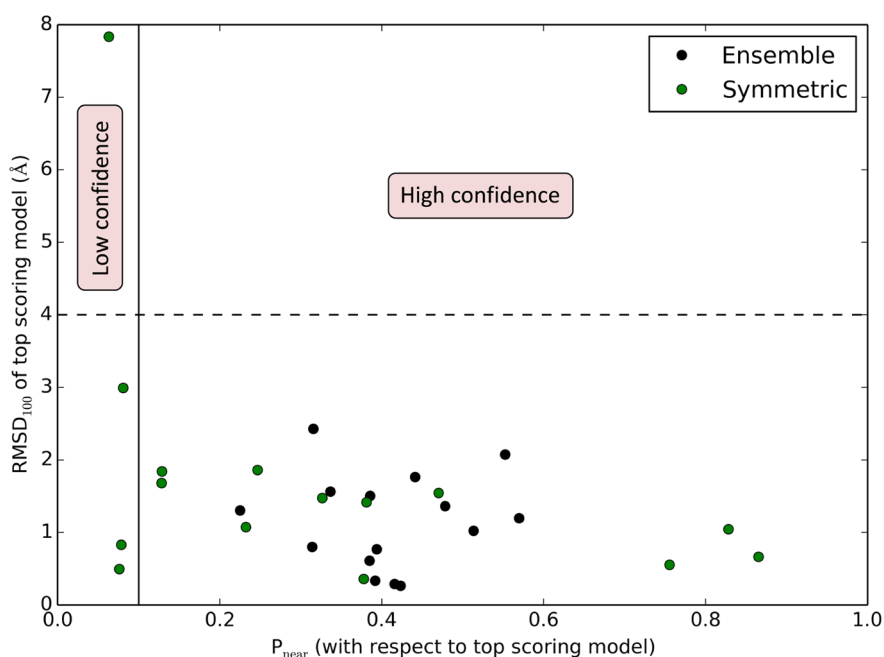
Figure 4. All atom, aligned $RMSD_{100}$ of the top scoring models as a function of the confidence metric (P_{near} with respect to the top scoring model) for both ensemble (black) and symmetric (green) docking. The solid vertical line ($P_{near} = 0.1$) indicates the separation between high (right) and low confidence (left). The dotted horizontal line indicates the cutoff for accurate predicted models (4 Å). All (26/26) high confidence structures were accurate. All (1/1) inaccurate predictions were low confidence.

Dependence of Prediction Accuracy on Input Structure Followed Expected Trends. Due to the difference in accuracy of the input structures (with respect to their structures in the complex), the overall docking results were expected to depend on the method used to obtain them. While Table S4 shows average metrics over all proteins for ensemble and symmetric docking, we also examined averages of each type of input structure (HM, unbound, and bound+perturbed) separately. The category-dependent results of f_{nat}, iRMSD, LRMSD, and DockQ score are shown in Figures 2B and 3B for ensemble and symmetric docking, respectively. First, similar to the total average results, the accuracy of the predicted structures was generally best when using all experimental data (cryo-EM and SID). When considering the difference between input structure categories, we expected the structures built from bound+perturbed to be the best followed by unbound and HM based on accuracy of input structures prior to docking. The results typically followed the expected trend for both ensemble and symmetric docking. The results were best for bound+perturbed, followed by unbound and HM, respectively. We also note that there were no cases where the inclusion of SID improved the prediction results for the bound+perturbed docking simulations (Isc+cryo-EM+SID compared to Isc+cryo-EM), which is indicated by the very similar results when comparing Isc+cryo-EM and Isc+cryo-EM+SID for bound+perturbed. However, this did not indicate poor performance from the SID score, rather this occurred because the scoring was already excellent. On the other hand, SID improvements were much more common for the docking simulations using homology models since the scoring was slightly off without SID. To summarize, whether SID can improve prediction results depends on the scoring before including SID (in that the SID score is most beneficial when the scoring was close to being accurate).

Confidence Metric Effectively Separated Nativelike and Non-Nativelike Structures. As not all predictions resulted in accurate structures ($RMSD_{100}$ less than 4 Å), we sought to quantify a prediction confidence metric that could be calculated without knowledge of the native structure. For the confidence metric (applied to both ensemble and symmetric docking), P_{near} of the score vs RMSD distribution with respect to the top scoring model was calculated. We hypothesized that when accurate predictions were made, a larger quantity of similar structures would score well and thus the funneling of the score vs RMSD distribution would be stronger (as quantified by a higher P_{near}).[22] Figure 4 shows the $RMSD_{100}$ of the top scoring model as a function of the confidence metric. Proteins for which the P_{near} was greater than 0.1 were identified as high confidence, as indicated by the vertical solid line. The prediction was accurate for all proteins identified as high confidence (26/26), as indicated by the dotted horizontal line. All inaccurate predictions were flagged as low confidence (1/1), with only three accurate predictions flagged as low confidence. All ensemble docking predictions (black, all accurate) were identified as high confidence (15/15). Overall, this confidence metric served as an effective indicator of prediction accuracy.

■ CONCLUSION

Surface-induced dissociation has typically been used on protein complexes to determine stoichiometry and intersubunit connectivity. We recently demonstrated that an appearance energy (AE) extracted from SID data also provides information on interface composition and developed models to predict AE from structure.[13] Using this predictive model, we showed that SID AE data were able to distinguish between nativelike and non-nativelike models of subcomplexes generated with rigid protein–protein docking in Rosetta using bound crystal structures as inputs. While previous work laid the foundation

of SID usage in protein complex structure prediction, there were a few shortcomings that we sought to address in this work, primarily to make the docking more realistic. The most notable shortcomings of previous work that were overcome here were the exclusive use of the exact tertiary structure from bound crystal structures, rigid docking, and only predicting full complexes in a few cases. To limit and frequently completely obviate required prior knowledge, we used three different types of input structures in the present work: homology models, unbound crystal structures, and bound+perturbed crystal structures. For each input structure, we generated an ensemble of backbones to encode flexibility. Additionally, we performed two types of docking: ensemble docking (with flexibility) and symmetric docking (which predicted full complexes in all cases). Finally, we included simulated low-resolution density maps from cryo-EM into scoring as well.

When ensemble docking was performed on all input structures, the $RMSD_{100}$ of the predicted structure without including experimental data was less than 4 Å for only 7/15 cases but was less than 4 Å in all (15/15) cases when both cryo-EM and SID data were included. Furthermore, when SID and cryo-EM data were included, all calculated CAPRI metrics improved on average. For the ensemble docking, there were both cryo-EM- and SID-driven cases. For the symmetric docking, the $RMSD_{100}$ of the predicted model was less than 4 Å for only 5/15 cases without including experimental data and also when cryo-EM density maps were incorporated. However, when SID and cryo-EM data were included, the $RMSD_{100}$ of the predicted model was less than 4 Å for 14/15 cases. Additionally, typically CAPRI metrics were the best when all experimental data were included.

In this study, we were able to overcome many shortcomings of previous docking studies with SID data. Full complex structures were built without the knowledge of the exact tertiary structure. To do this, we combined data in the forms of SID AE and noisy, low-resolution cryo-EM density maps to build full structures of protein complexes. This work demonstrates the potential of SID as a tool to facilitate the modeling of protein complexes. We hypothesized that SID AE and low-resolution cryo-EM density maps would provide complementary information for the scoring of complexes. The experimental methods are also compatible, with cryo-EM becoming more accessible as it grows in popularity and with SID commercialization underway.

However, in this study, we generally observed that scoring with cryo-EM may be too accurate to reap the full benefits of scoring with SID. While the final prediction results were excellent, we only observed a few cases where the inclusion of SID appeared strongly beneficial if cryo-EM data were already applied (Isc+cryo-EM+SID over Isc+cryo-EM). However, since we have consistently observed that the SID scoring function scores most accurate structures favorably and penalizes most inaccurate structures (see Figure S2), we believe that the lack of consistent SID benefit here was simply because of the cryo-EM accuracy, even for low-resolution density maps. For this reason, the results of this work motivate the potential of modeling with SID and data that provide less structural information than cryo-EM density maps. There are multiple examples of MS methods that may fit this category, namely ion mobility, covalent labeling, and chemical cross-linking and non-MS methods such as SAXS. We hope to find the optimal amount of additional information to include with SID for the best results in these realistic docking simulations.

While we hoped that SID and cryo-EM would be the most compatible, the MS methods are easier to perform and thus offer even more potential for experimental ease. Based on the scoring strategy presented here, different types of experimental data can be easily combined with SID. Ultimately, future work will focus on determining the optimal use of SID data with other methods that provide additional orthogonal structural information.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c05468.

> Additional methodological details; Table S1, list of interfaces for ensemble docking; Table S2, list of symmetries for symmetric docking; Table S3, relative weights used for scoring; Table S4, averages of metrics; Figure S1, frequency of RMSD from backbone sampling; Figure S2, SID score vs RMSD plots for 1SWB; Figure S3, symmetric docked structures for 8TIM; and tutorial outlining methods used to perform docking and scoring (PDF)

■ 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

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

Shane M. Canfield − *Department of Chemistry, Kenyon College, Gambier, Ohio 43022, United States*

Sophie R. Harvey − *Department of Chemistry and Biochemistry, 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;* ⓘ orcid.org/0000-0003-0495-2538

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.analchem.0c05468

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Leelananda, S. P.; Lindert, S. *Beilstein J. Org. Chem.* **2016**, *12*, 2694−2718.

(2) Shen, H. B.; Chou, K. C. *J. Proteome Res.* **2009**, *8* (3), 1577−84.

(3) Liko, I.; Allison, T. M.; Hopper, J. T.; Robinson, C. V. *Curr. Opin. Struct. Biol.* **2016**, *40*, 136−144.

(4) Ben-Nissan, G.; Sharon, M. *Curr. Opin. Chem. Biol.* **2018**, *42*, 25−33.

(5) van de Waterbeemd, M.; Fort, K. L.; Boll, D.; Reinhardt-Szyba, M.; Routh, A.; Makarov, A.; Heck, A. J. *Nat. Methods* **2017**, *14* (3), 283−286.

(6) Blackwell, A. E.; Dodds, E. D.; Bandarian, V.; Wysocki, V. H. *Anal. Chem.* **2011**, *83* (8), 2862−5.

(7) Zhou, M.; Jones, C. M.; Wysocki, V. H. *Anal. Chem.* **2013**, *85* (17), 8262−7.

(8) Harvey, S. R.; Yan, J.; Brown, J. M.; Hoyes, E.; Wysocki, V. H. *Anal. Chem.* **2016**, *88* (2), 1218−21.

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

(10) Sahasrabuddhe, A.; Hsia, Y.; Busch, F.; Sheffler, W.; King, N. P.; Baker, D.; Wysocki, V. H. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (6), 1268−1273.

(11) Song, Y.; Nelp, M. T.; Bandarian, V.; Wysocki, V. H. *ACS Cent. Sci.* **2015**, *1* (9), 477−487.

(12) 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.

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

(14) Hofmann, T.; Fischer, A. W.; Meiler, J.; Kalkhof, S. *Methods* **2015**, *89*, 79−90.

(15) Kahraman, A.; Herzog, F.; Leitner, A.; Rosenberger, G.; Aebersold, R.; Malmstrom, L. *PLoS One* **2013**, *8* (9), No. e73411.

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

(17) Walzthoeni, T.; Joachimiak, L. A.; Rosenberger, G.; Rost, H. L.; Malmstrom, L.; Leitner, A.; Frydman, J.; Aebersold, R. *Nat. Methods* **2015**, *12* (12), 1185−90.

(18) Bullock, J. M. A.; Schwab, J.; Thalassinos, K.; Topf, M. *Mol. Cell Proteomics* **2016**, *15* (7), 2491−500.

(19) Mendoza, V. L.; Vachet, R. W. *Mass Spectrom. Rev.* **2009**, *28* (5), 785−815.

(20) Roberts, V. A.; Pique, M. E.; Hsu, S.; Li, S. *Biochemistry* **2017**, *56* (48), 6329−6342.

(21) Xie, B.; Sood, A.; Woods, R. J.; Sharp, J. S. *Sci. Rep* **2017**, *7* (1), 4552.

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

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

(24) Biehn, S. H.; Lindert, S. *Accurate Protein Structure Prediction with Hydroxyl Radical Footprinting Data* **2021**, *12*, 341.

(25) Marklund, E. G.; Degiacomi, M. T.; Robinson, C. V.; Baldwin, A. J.; Benesch, J. L. *Structure* **2015**, *23* (4), 791−9.

(26) Politis, A.; Park, A. Y.; Hall, Z.; Ruotolo, B. T.; Robinson, C. V. *J. Mol. Biol.* **2013**, *425* (23), 4790−801.

(27) Degiacomi, M. T. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (1), 113−117.

(28) Eschweiler, J. D.; Frank, A. T.; Ruotolo, B. T. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (10), 1991−2000.

(29) Robertson, J. C.; Nassar, R.; Liu, C.; Brini, E.; Dill, K. A.; Perez, A. *Proteins: Struct., Funct., Genet.* **2019**, *87* (12), 1333−1340.

(30) Weiner, B. E.; Alexander, N.; Akin, L. R.; Woetzel, N.; Karakas, M.; Meiler, J. *Proteins: Struct., Funct., Genet.* **2014**, *82* (4), 587−95.

(31) Vernon, R.; Shen, Y.; Baker, D.; Lange, O. F. *J. Biomol. NMR* **2013**, *57* (2), 117−27.

(32) Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. *Methods* **2009**, *49* (2), 174−80.

(33) Lindert, S.; Staritzbichler, R.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. *Structure* **2009**, *17* (7), 990−1003.

(34) Lindert, S.; Alexander, N.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. *Structure* **2012**, *20* (3), 464−78.

(35) Lindert, S.; Hofmann, T.; Wötzel, N.; Karakaş, M.; Stewart, P. L.; Meiler, J. *Biopolymers* **2012**, *97* (9), 669−77.

(36) Frenz, B.; Walls, A. C.; Egelman, E. H.; Veesler, D.; DiMaio, F. *Nat. Methods* **2017**, *14* (8), 797−800.

(37) Schindler, C. E. M.; de Vries, S. J.; Sasse, A.; Zacharias, M. *Structure* **2016**, *24* (8), 1387−1397.

(38) Ignatov, M.; Kazennov, A.; Kozakov, D. *J. Mol. Biol.* **2018**, *430* (15), 2249−2255.

(39) Jiménez-García, B.; Pons, C.; Svergun, D. I.; Bernadó, P.; Fernández-Recio, J. *Nucleic Acids Res.* **2015**, *43* (W1), W356−61.

(40) MacCallum, J. L.; Perez, A.; Dill, K. A. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (22), 6985−90.

(41) Del Alamo, D.; Tessmer, M. H.; Stein, R. A.; Feix, J. B.; McHaourab, H. S.; Meiler, J. *Biophys. J.* **2020**, *118* (2), 366−375.

(42) Bonomi, M.; Pellarin, R.; Kim, S. J.; Russel, D.; Sundin, B. A.; Riffle, M.; Jaschob, D.; Ramsden, R.; Davis, T. N.; Muller, E. G.; Sali, A. *Mol. Cell Proteomics* **2014**, *13* (11), 2812−23.

(43) Seffernick, J. T.; Lindert, S. *J. Chem. Phys.* **2020**, *153* (24), 240901.

(44) Marzolf, D. R.; Seffernick, J. T.; Lindert, S. *J. Chem. Theory Comput* **2021**, *17*, 2619.

(45) Boyken, S. E.; Benhaim, M. A.; Busch, F.; Jia, M.; Bick, M. J.; Choi, H.; Klima, J. C.; Chen, Z.; Walkey, C.; Mileant, A.; Sahasrabuddhe, A.; Wei, K. Y.; Hodge, E. A.; Byron, S.; Quijano-Rubio, A.; Sankaran, B.; King, N. P.; Lippincott-Schwartz, J.; Wysocki, V. H.; Lee, K. K.; Baker, D. *Science* **2019**, *364* (6441), 658−664.

(46) Chen, Z.; Kibler, R. D.; Hunt, A.; Busch, F.; Pearl, J.; Jia, M.; VanAernum, Z. L.; Wicky, B. I. M.; Dods, G.; Liao, H.; Wilken, M. S.; Ciarlo, C.; Green, S.; El-Samad, H.; Stamatoyannopoulos, J.; Wysocki, V. H.; Jewett, M. C.; Boyken, S. E.; Baker, D. *Science* **2020**, *368* (6486), 78−84.

(47) Zhou, M.; Huang, C.; Wysocki, V. H. *Anal. Chem.* **2012**, *84* (14), 6016−23.

(48) Yan, J.; Zhou, M.; Gilbert, J. D.; Wolff, J. J.; Somogyi, Á.; Pedder, R. E.; Quintyn, R. S.; Morrison, L. J.; Easterling, M. L.; Paša-Tolić, L.; Wysocki, V. H. *Anal. Chem.* **2017**, *89* (1), 895−901.

(49) VanAernum, Z. L.; Gilbert, J. D.; Belov, M. E.; Makarov, A. A.; Horning, S. R.; Wysocki, V. H. *Anal. Chem.* **2019**, *91* (5), 3611−3618.

(50) Snyder, D. T.; Panczyk, E. M.; Somogyi, A.; Kaplan, D. A.; Wysocki, V. *Anal. Chem.* **2020**, *92* (16), 11195−11203.

(51) Andre, I.; Bradley, P.; Wang, C.; Baker, D. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (45), 17656−61.

(52) Chaudhury, S.; Berrondo, M.; Weitzner, B. D.; Muthu, P.; Bergman, H.; Gray, J. J. *PLoS One* **2011**, *6* (8), e22477−e22477.

(53) DiMaio, F.; Song, Y.; Li, X.; Brunner, M. J.; Xu, C.; Conticello, V.; Egelman, E.; Marlovits, T.; Cheng, Y.; Baker, D. *Nat. Methods* **2015**, *12* (4), 361−365.

(54) Wang, R. Y.; Song, Y.; Barad, B. A.; Cheng, Y.; Fraser, J. S.; DiMaio, F. *Elife* **2016**, *5*, e17219.

(55) Basu, S.; Wallner, B. *PLoS One* **2016**, *11* (8), No. e0161879.

(56) Ohio Supercomputer Center. Ohio Supercomputer Center: 1987.