Predicting Protein Complex Structure from Surface Induced Dissociation Mass Spectrometry Data

Justin T. Seffernick,[1] Sophie R. Harvey, [1] Vicki H. Wysocki, [1] and Steffen Lindert[1,*]

[1]Department of Chemistry and Biochemistry and Resource for Native Mass Spectrometry Guided Structural Biology, Ohio State University, Columbus, OH, 43210

* Correspondence to:
Department of Chemistry and Biochemistry, Ohio State University
2114 Newman & Wolfrom Laboratory, 100 W. 18th Avenue, Columbus, OH 43210
614-292-8284 (office), 614-292-1685 (fax)
lindert.1@osu.edu

Methods

SID Dataset

The dataset used for evaluation of the protein-protein docking poses using SID data consisted of a homodimer (triose phosphate isomerase, 8tim, 11+ charge state), a homotetramer (streptavidin, 1swb,[1] 11+), a heterotetramer (hemoglobin, 1gzx,[2] 11+), three homopentamers (cholera toxin B, 1fgb,[3] 11+; C-reactive protein, 1gnh,[4] 17+; serum amyloid P, 1sac,[5] 19+), and a homohexamer (glutamate dehydrogenase, 3mvo,[6] 25+). Medium- to high-resolution (< 3.5 Å resolution) crystal structures of the intact protein complexes existed for all members of the dataset and were used for evaluation purposes.

Triose phosphate isomerase, hemoglobin, cholera toxin B, and glutamate dehydrogenase were purchased from Sigma-Aldrich (St. Louis, MO, USA). Serum Amyloid P and c-reactive protein were purchased from CalBioChem (EMD Biosciences, Inc., San Diego, CA, USA) and streptavidin was purchased from Thermo Scientific Pierce Biotechnology (Rockford, IL, USA). All samples were analyzed at approximately 10 μM complex concentration in 80 mM ammonium acetate (Sigma Aldrich, St. Louis, MO, USA) plus 20 mM triethyl ammonium acetate (TEAA) (Sigma Aldrich, St. Louis, MO, USA). We used TEAA to produce 'charge reducing' conditions, which are thought to keep the complex more compact and native-like.[7-9] Residual salt impurities were removed by buffer exchanging using micro Spin 6 columns (Bio-Rad, Hercules, CA, USA), as required.

All SID experiments were performed on in-house modified Synapt G2 or G2S instruments (Waters, Milford, UK). The instruments were modified as previously described,[10] however, in this case the SID device was placed between the Trap ion guide/collision cell and the IM cell and hence referred to as Trap-SID. All proteins were introduced into the mass spectrometer using nano-electrospray ionization in positive mode. Nano-electrospray tips were made in-house using thin-walled glass capillaries (i.d. 0.8 mm) using a Flaming/Brown micropipette puller (Sutter Instrument Company, Novato, CA, USA). The spray voltage (typically 1-1.4 kV) was applied using a thin (0.368 mm) platinum wire (Alfa Aesar, Ward Hill, MA, USA). The instrument was operated in ion mobility, sensitivity mode. For MS experiments, a Trap gas flow rate of 4 mL/min and a Trap DC bias of 45 V are applied and the SID device tuned to give a 1-5 V difference between the Trap exit and the entrance lens of the SID device and a 5-10 V difference between the exit lens of the SID device and the helium cell entrance. For SID experiments, the SID device is tuned to steer the ions for collision with the surface. The SID acceleration voltage is defined by

the potential difference between the DC offset of the Trap and the surface and can be adjusted using the Trap bias setting. The collision energy is determined by multiplying the collision voltage by the charge state of the precursor. For SID experiments, the Trap gas flow rate was lowered to 2 mL/min, to limit any gas-collisions which would cause CID. For SID, typically the most intense, unique, mass-to-charge species in the charge-reducing conditions was chosen for study (see above for charge states used). Performing SID over a range of different collision energies and determining the relative abundance of precursor and products at each energy allows energy-resolved mass spectrometry plots to be produced and used for AE estimation. For each protein, ERMS plots were used to identify which interfaces fragmented during SID and to determine the experimental appearance energies, arbitrarily defined as the acceleration energy needed to reach 10% intensity with respect to the intensity of the native complex (to avoid the influence of hot/pre-fragmenting precursor ions). Appearance energies were subsequently normalized by the number of inter-subunit protein-protein contacts in order to properly account for the non-interface-dependent rigidity factor (RF, see main text).

Table S1: Complexes used for docking using ideal (computationally predicted from crystal structures) SID AE data. Percent helix and strand values were calculated from relaxed crystal structures using DSSP.[11]

PDB ID	Complex type	Number of residues (per subunit)	Percent Helix	Percent Strand
1eym	Homodimer	107	10.6	39.8
1f37	Homodimer	110	29.4	20.8
1ix9	Homodimer	205	57.3	11.7
1qlw	Homodimer	328	33.6	19.3
1x8j	Homodimer	351	50.1	9.4
2car	Homodimer	196	36.5	24.7
2qcq	Homodimer	110	17.1	41.5
2vha	Homodimer	287	38.2	24.4
2voc	Homodimer	112	31.0	24.4
2xdi	Homodimer	107	59.9	0.0
3cby	Homodimer	108	18.1	35.8
3cdy	Homodimer	109	2.8	49.5
3e18	Homodimer	359	35.5	23.0
3fl1	Homodimer	252	45.2	14.1
3gmx	Homodimer	154	19.0	32.9
3hg5	Homodimer	398	28.7	23.5
3o1n	Homodimer	276	45.5	20.0
3vm9	Homodimer	153	79.2	0.0
4amb	Homodimer	400	42.6	14.8
4iwh	Homodimer	363	43.1	21.3
4r8d	Homodimer	394	42.4	15.1
4u13	Homodimer	109	31.8	44.5
4unu	Homodimer	111	2.7	48.6
5fi3	Homodimer	357	30.4	29.1
5idb	Homodimer	142	0.0	65.0

5j4g	Homodimer	107	0.0	59.6
5k1l	Homodimer	137	77.2	1.4
5onc	Homodimer	399	37.4	22.4
5yok	Homodimer	100	4.0	52.5
6ahp	Homodimer	110	45.0	31.1
3n9g	Heterodimer	225	5.1	47.8
4hpj	Heterodimer	332.5	46.0	17.9
4nzu	Heterodimer	218	5.7	49.4
4o5l	Heterodimer	235	7.1	47.6
1di0	Homopentamer	158	53.7	19.8
1nlq	Homopentamer	108	0.0	59.0
2x00	Homopentamer	228	10.4	50.8
3ck6	Homopentamer	252	51.6	15.9
3jcf	Homopentamer	351	54.3	14.0
3wtl	Homopentamer	214	8.9	48.3
4afh	Homopentamer	230	11.7	47.9
4avs	Homopentamer	204	6.1	45.4
4b5d	Homopentamer	230	12.7	47.4
4kly	Homopentamer	259	83.2	0.0
4x17	Homopentamer	272	6.2	40.1
5h5t	Homopentamer	204	10.4	49.1
5jrw	Homopentamer	373	57.3	14.2
5l4e	Homopentamer	317	35.7	14.2
5lzg	Homopentamer	103	24.0	38.0
5t2k	Homopentamer	248	37.5	31.8
6hin	Homopentamer	450	36.2	28.3
6qb7	Homopentamer	163	18.1	43.1
1gc0	Homotetramer	398	40.7	16.2
3e6g	Homotetramer	400	39.9	16.6
3wcc	Homotetramer	365	73.2	0.0
4hy3	Homotetramer	365	38.3	15.1
4ix2	Homotetramer	366	34.1	22.2

Table S2: Average RMSD's of top 100 scoring models with Rosetta and Rosetta with SID. Results show improvement in 8/9 cases.

Protein	Average RMSD of top 100 for Rosetta (Å)	Average RMSD of top 100 for Rosetta with SID (Å)
1fgb	15.44	14.74
1gnh	21.01	20.94
1gzx	21.61	18.20
1sac	23.20	21.04
1swb	17.91	15.72

3mvo	33.96	26.56
8tim	26.83	27.30
1gzx_dimers	29.51	27.24
1swb_dimers	24.13	24.13

Table S3: P_{near} values for Rosetta and Rosetta with SID to quantify funneling of score vs. RMSD plots. Three cases (1gzx, 1sac, 1gzx_dimers) show a drastic improvement (>3-fold).

Protein	P_{near} Rosetta	P_{near} Rosetta with SID	X-fold increase
1fgb	0.757	0.562	0.742
1gnh	0.00504	0.00522	1.04
1gzx	0.0115	0.485	42.2
1sac	0.0145	0.0534	3.68
1swb	0.945	0.964	1.02
3mvo	1.77e-14	3.70e-30	2.09e-16
8tim	0.000106	2.82e-12	2.70e-8
1gzx_dimers	3.66e-5	0.00155	42.4
1swb_dimers	0.0171	0.0171	1.00

Docking Tutorial

To use RosettaDock to rescore structures with SID AE, two main stages need to be performed in Rosetta:

1. Generate docked structures using RosettaDock.
2. Rescore the poses using Rosetta SID_rescore application.

In this tutorial, variables that need to be specified by the user are shown in brackets (< >). Dockings and analysis from this paper were performed using talaris2014 scoring function. To use the scoring function, include the flag `-restore_talaris_behavior` in all Rosetta command lines. Without this flag, REF15 will be used by default.

Step 1: RosettaDock

a. Prepare a pdb file containing both partners in a predicted starting position.
b. Prepack the chains by running the following command:
   `~/Rosetta/main/source/bin/docking_prepack_protocol.default.<os><compiler>release -in:file:s <pdb> -partners <chains>`
   - <os> operating system (macos, linux).
   - <compiler> compiler used (gcc, clang, etc.).
   - <pdb> name of coordinate file in pdb format.
   - <chains> chains of docking partners, separated by underscore. ex: A_B where A is the static chain, B is the mobile chain.

   Example: If gcc was used to compile on linux to dock chains A and C of complex.pdb:
   `~/Rosetta/main/source/bin/docking_prepack_protocol.default.linuxgccrelease -in:file:s complex.pdb -partners A_C`
c. To dock the chains, use the following command:
   `~/Rosetta/main/source/bin/docking_protocol.default.<os><compiler>release -in:file:s <prepacked_pdb> -partners <chains> -nstruct <n_structs>`

- <prepacked_pdb> output pdb from prepack step.
- <n_structs> number of structures to generate (>10,000 recommended)
- Additionally, a randomization flag (-randomize1, -randomize2, or -spin) can be given to search more conformational space.

Example: Dock 10,000 structures of chains A and C of complex_0001.pdb

```
~/Rosetta/main/source/bin/docking_protocol.default.linuxgccrelease -in:file:s complex_0001.pdb -partners A_C -nstruct 10000
```

Step 1 will result in <n_structs> docked structures. Step 2 will use SID data to rescore and rank the generated structures. A more detailed description for generating docked structures can be found here: https://www.rosettacommons.org/demos/latest/tutorials/Protein-Protein-Docking/Protein-Protein-Docking

Step 2: SID_rescore

a. First create a text file containing the names of the docked structures (pdb files).
   Example:

```
complex_0001_00001.pdb
complex_0001_00002.pdb
complex_0001_00003.pdb
complex_0001_00004.pdb
complex_0001_00005.pdb
…
complex_0001_10000.pdb
```

b. To rescore the poses, run the SID_rescore application using the following command:

```
~/Rosetta/main/source/bin/SID_rescore.default.<os><compiler>release -in:file:l <file_with_docked_poses> -AE <AE_from_SID> -interface <chains> -n_ints <n_ints> -out:file:o <output_file> -native <native_pdb>
```

   - <file_with_docked_poses> file created in Step 2a.
   - <AE_from_SID> appearance energy from SID experiment (eV).
   - <n_ints> number of intra-chain contacts in docking.
   - <output_file> (*optional*) name of output file from this command. (SID_rescore_default.out by default)
   - <native_pdb> (*optional*) native pdb. Will calculate RMSD if given.

   Example: Rescoring list_docked_pdbs.txt (list of pdbs). Experimental AE of 100.0 eV, one intra-chain contact, output file named complex_docking_scores.out, and calculate RMSD to native.pdb.

```
~/Rosetta/main/source/bin/SID_rescore.default.linuxgccrelease -in:file:l list_docked_pdbs.txt -AE 100.0 -interface A_C -n_ints 1 -out:file:o complex_docking_scores.out -native native.pdb
```

Step 2 will result in an output file containing the predicted AE, Rosetta_score, SID_score, Rosetta_SID_score, and RMSD (if native specified) for each of the docked structures. Use the Rosetta_SID_score value to sort the poses and thus select optimal predicted structures.
Example output file:

```
Pose_number AE_pred     Rosetta_score     SID_score   Rosetta_SID_score RMSD
1           7.651       -87.837           0.00703     -87.795           18.34
2           44.428      -91.943           0.0         -91.943           2.053
…
```

10000	-18.96	-87.997	0.133	-87.197	15.33

Figure S1: Docked complexes of five subcomplexes for which including SID restraints (from ideal AE data) improved the RMSD by more than 14 Å (3vm9, 3gmx, 3jcf, 4ix2, and 4hy3). Green structures are the natives, blue are the models predicted without SID data, and red are models predicted with the Bayesian Rosetta SID rescore. For each dimer, the stationary subunit (left) was aligned to show the discrepancy or lack thereof for the mobile (docked) subunit (right).

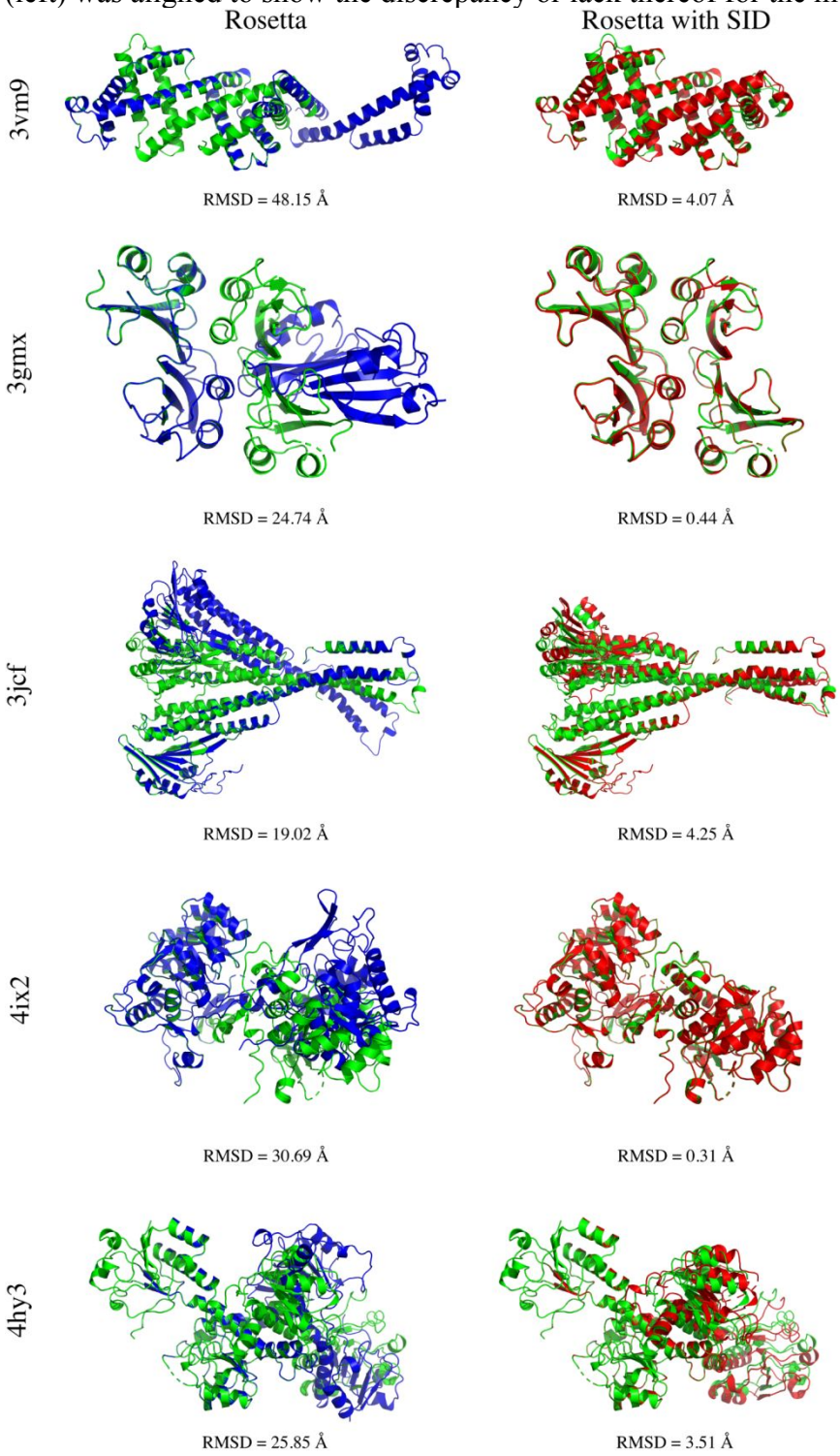


Figure S2: Raw SID Score vs. RMSD plots for 1gzx, 1sac, 1swb, and 1gzx_dimers. SID score generally scored low RMSD models well while penalizing most high-RMSD structures.

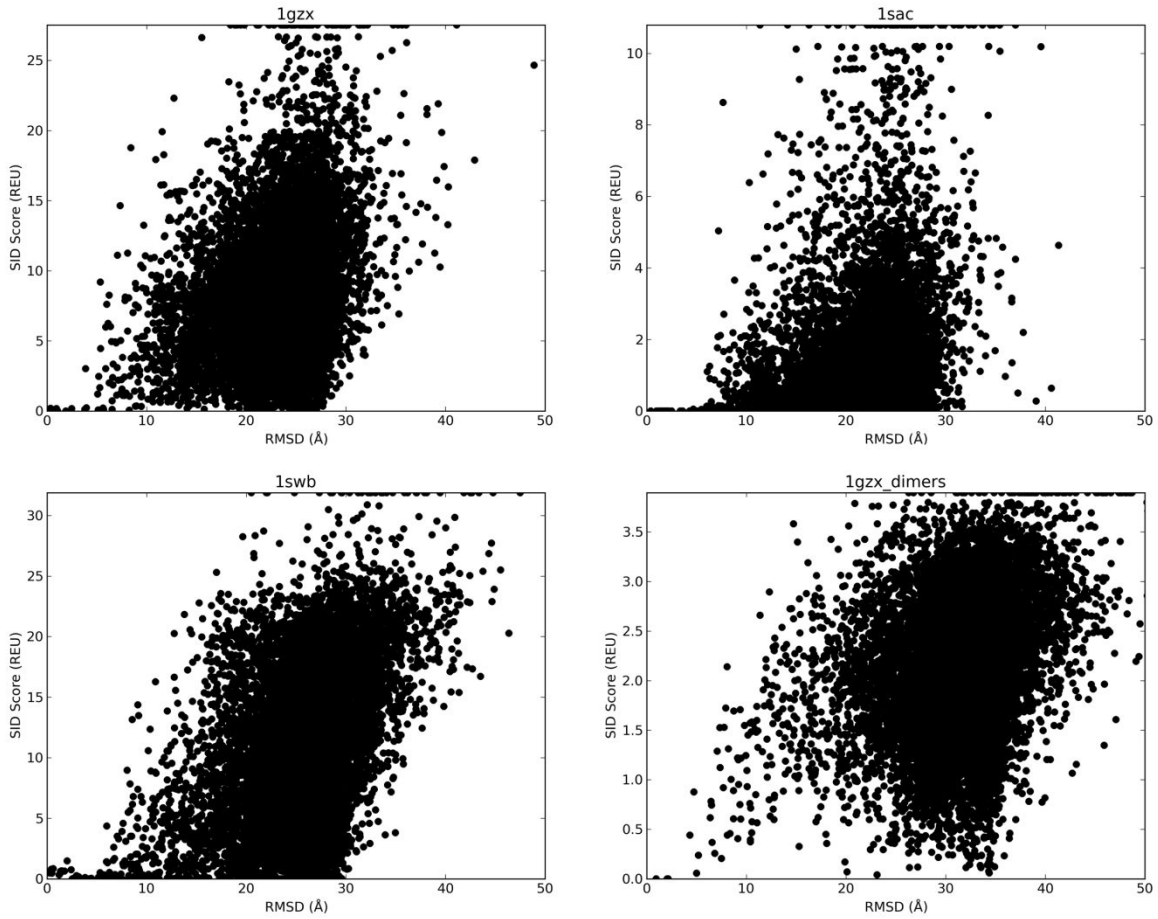


Figure S3: Comparison of funneling metrics with the use of ideal AE (predicted from crystal structures): P_{near} (A) and score difference between high RMSD models and minimum score (B). P_{near} improved for 56/57 cases when SID ideal AE was used and average high RMSD separation improved for all cases when ideal SID was used.

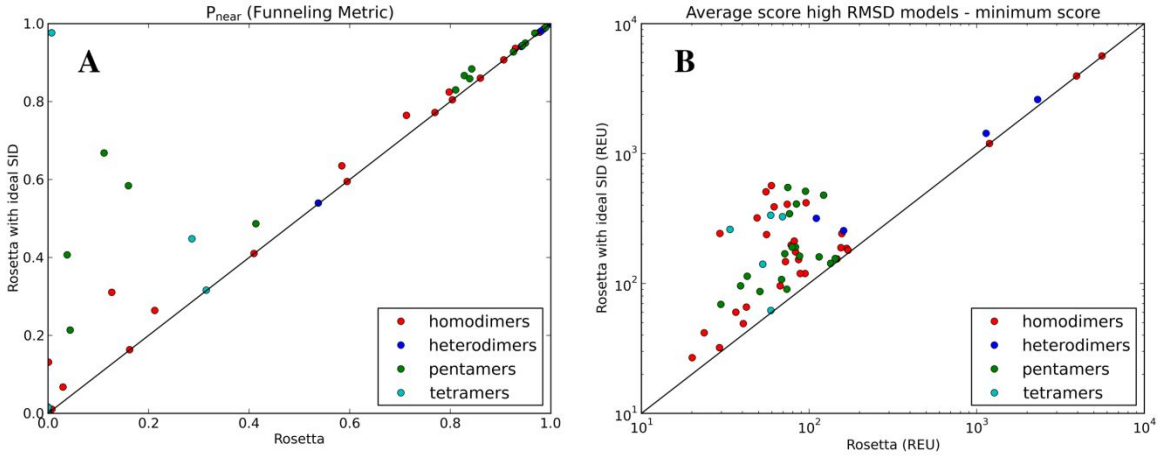


Figure S4: Score vs. RMSD plots of each complex for which P_{near} (quantification of "goodness of funneling") decreased by more than half (absolute values in Table S2) when SID data was used. 8tim: 2.70e-8-fold increase, 3mvo: 2.09e-16-fold increase.

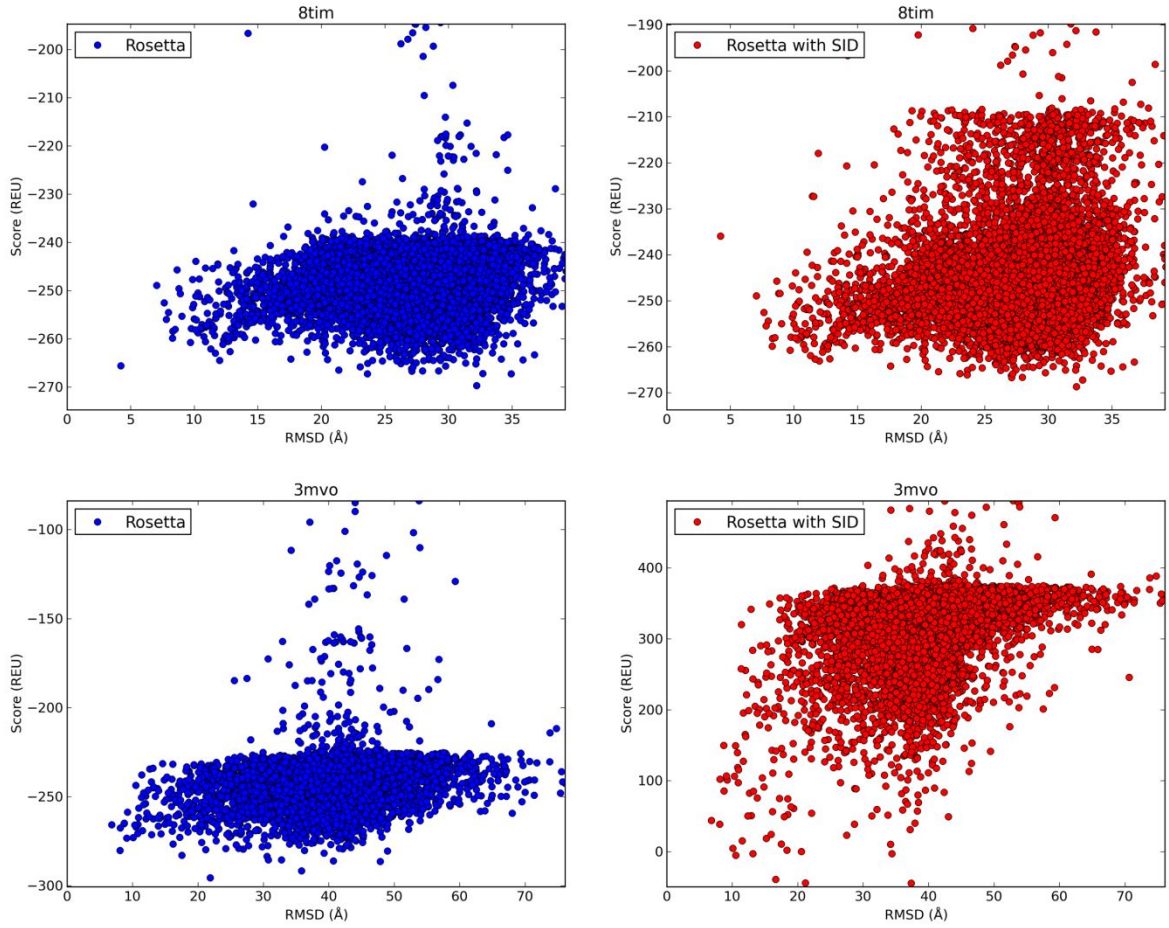
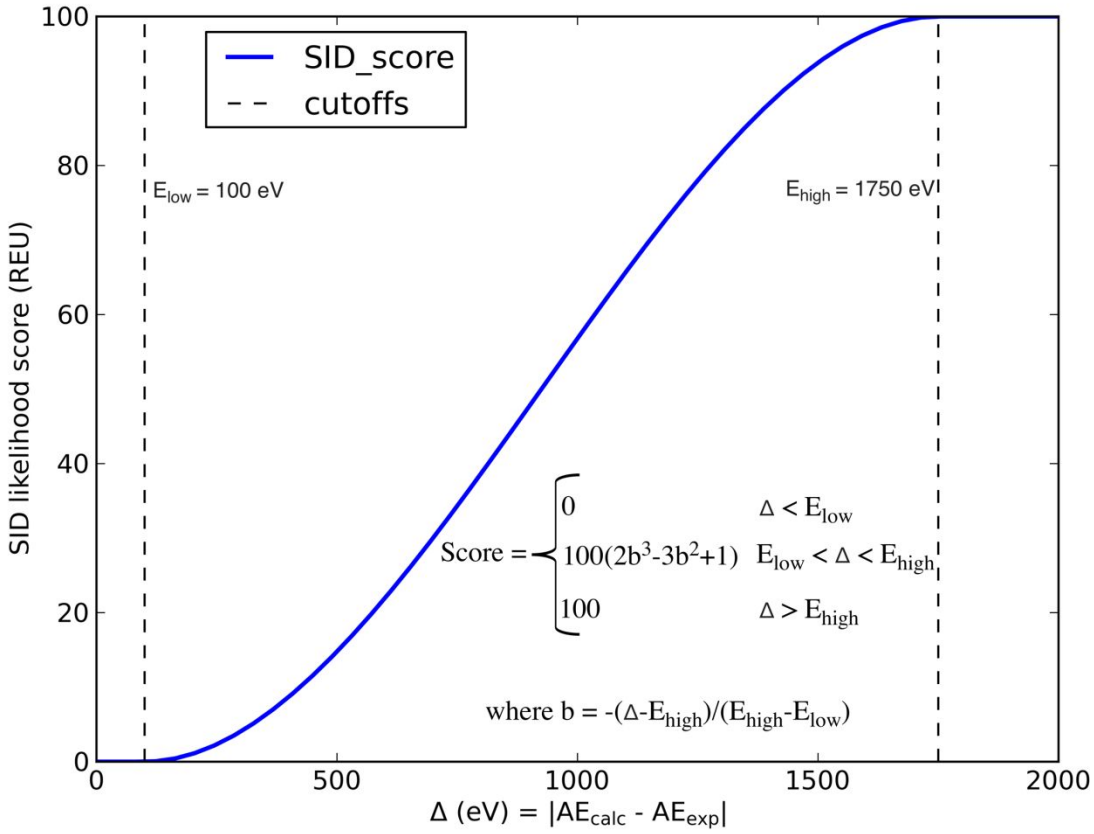


Figure S5: Function used to evaluate SID likelihood score. Structures with small deviation from the measured experimental data (low Δ) have lower scores and thus higher probability while structures with large deviation from the measured experimental data (high Δ) have higher scores and thus lower probability. This function contains two cutoffs, a lower cutoff (E_{low}, below which the score is minimum) and a higher cutoff (E_{high}, above which the score is maximum). We hypothesize that the inclusion of the E_{low} helps account for experimental error.



References:

1. Freitag, S.; Trong, I. L.; Klumb, L.; Stayton, P. S.; Stenkamp, R. E., Structural studies of streptavidin binding loop. Protein Sci: 1997; Vol. 6, pp 1157-1166.
2. Paoli, M.; Liddington, R.; Tame, J.; Wilkinson, A.; Dodson, G., Crystal Structure of T State Haemoglobin with Oxygen Bound At All Four Haems. J. Mol. Biol.: 1996; Vol. 256, pp 775-792.
3. Zhang, R.-G.; Westbrook, M. L.; Westbrook, E. M.; Scott, D. L.; Otwinowski, Z.; Maulik, P. R.; Reed, R. A.; Shipley, G. G., The 2.4 Å Crystal Structure of Cholera Toxin B Subunit Pentamer: Choleragenoid. J. Mol. Biol.: 1995; Vol. 251, pp 550-562.
4. Shrive, A. K.; Cheetham, G. M. T.; Holden, D.; Myles, D. A. A.; Turnell, W. G.; Volanakis, J. E.; Pepys, M. B.; Bloomer, A. C.; Greenhough, T. J., Three dimensional structure of human C-reactive protein. Nat. Struct. Mol. Biol.: 1996; Vol. 3, pp 346-354.
5. Emsley, J.; White, H. E.; O'Hara, B. P.; Olivia, G.; Srnivasan, N.; Tickle, I. J.; Blundell, T. L.; Pepys, M. B.; Wood, S. P., Structure of pentameric human serum amyloid P component. Nature: 1994; Vol. 367, pp 338-367.
6. Bailey, J.; Powell, L.; Sinanan, L.; Neal, J.; Li, M.; Smith, T.; Bell, E., A Novel Mechanism of V Type Zinc Inhibition of Glutamate Dehydrogenase Results from Disruption of Subunit Interactions Necessary for Efficient Catalysis. FEBS J.: 2011; Vol. 278, pp 3140-3151.
7. Zhou, M.; Dagan, S.; Wysocki, V. H., Impact of charge state on gas-phase behaviors of noncovalent protein complexes in collision induced dissociation and surface induced dissociation. Analyst: 2013; Vol. 138, pp 1353-1362.
8. Pagel, K.; Hyung, S.-J.; Ruotolo, B. T.; Robinson, C. V., Alternate Dissociation Pathways Identified in Charge-Reduced Protein Complex Ions. Anal. Chem.: 2010; Vol. 82, pp 5363-5372.
9. Hall, Z.; Politis, A.; Bush, M. F.; Smith, L. J.; Robinson, C. V., Charge-State Dependent Compaction and Dissociation of Protein Complexes: Insights from Ion Mobility and Molecular Dynamics J. Am. Chem. Soc.: 2012; Vol. 134, pp 3429-3438.
10. Zhou, M.; Huang, C.; Wysocki, V. H., Surface-Induced Dissociation of Ion Mobility-Separated Noncovalent Complexes in Quadrupole/Time-of-Flight Mass Spectrometer. *Anal. Chem.*: **2012**; Vol. 84, pp 6016-6023.
11. Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G., A series of PDB-related databanks for everyday needs. Nucleic Acids Res: 2014; Vol. 43, pp D364-D368.