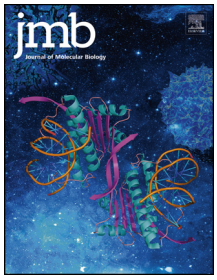

A Membrane Protein Complex Docking Benchmark

Panagiotis I. Koukos, Inge Faro†, Charlotte W. van Noort† and Alexandre M.J.J. Bonvin

Bijvoet Center for Biomolecular Research, *Faculty of Science—Chemistry, Utrecht University, Padualaan 8, Utrecht 3584CH, the Netherlands*

Correspondence to Alexandre M.J.J. Bonvin: *a.m.j.j.bonvin@uu.nl*
https://doi.org/10.1016/j.jmb.2018.11.005
Edited by Michael Sternberg

Abstract

We report the first membrane protein–protein docking benchmark consisting of 37 targets of diverse functions and folds. The structures were chosen based on a set of parameters such as the availability of unbound structures, the modeling difficulty and their uniqueness. They have been cleaned and consistently numbered to facilitate their use in docking. Using this benchmark, we establish the baseline performance of HADDOCK, without any specific optimization for membrane proteins, for two scenarios: true interface-driven docking and *ab initio* docking. Despite the fact that HADDOCK has been developed for soluble complexes, it shows promising docking performance for membrane systems, but there is clearly room for further optimization. The resulting set of docking decoys, together with analysis scripts, is made freely available. These can serve as a basis for the optimization of membrane complex-specific scoring functions.

Introduction

The docking community makes extensive use of benchmarks for evaluating the performance of docking algorithms and constantly improving them. Such benchmarks are also critical to allow a fair comparison between various algorithms, next to blind docking experiments such as CAPRI [1] for protein–protein and protein–peptide docking and Drug Design Data Resource grand challenges (D3R) [2,3] for small-molecule docking. Some of the most cited benchmarks are the protein–protein [4], protein–peptide [5,6], protein–DNA [7] and protein–ligand [8] ones. Several recent publications have made use of membrane protein-related benchmarks for testing and validating their software. RosettaMP [9], a recently updated addition to Rosetta's toolbox, supports a general membrane representation and can be used in combination with many of Rosetta's existing sampling and scoring protocols. The MPdock protocol is a combination of the RosettaMP and RosettaDock protocols [10] and supports docking of membrane proteins. The same publication also presents the MPsymdock protocol, which can be used to assemble homomeric membrane protein complexes from their monomeric constituents using known symmetry information. The authors also tested the newly minted protocols on membrane protein complexes; however, since they intended those demonstrations as a proof of concept, they only tested on five and four complexes for the MPdock and MPsymdock protocols, respectively. Other researchers have made use of more extensive data sets for their work, such as the Memdock [11] software and the modification to the scoring schemes employed by DOCK/PIERR [12,13]. In the case of the former, their training and testing sets consisted of 43 and 21 complexes (for a total of 64) obtained from the OPM database [14]; however, all entries are helical proteins. The same is true for DOCK/PIERR as well. Their data set makes use of the Membrane Proteins of Known Structure (MPSTRUC) database as the primary source of data and contains 22 biological complexes as well 8 artificial complexes, which have been created by separating GPCR proteins into separate parts after cutting them at one of the cytosolic/extracellular loops. This data set mostly consists of GPCRs and small helical complexes. None of the aforementioned works make the structures they used available, and

therefore, their data sets cannot be used as a docking benchmark. Another GPCR data set has been recently published [15]. To the best of our knowledge, however, there is no general and non-redundant docking benchmark for membrane protein–protein complexes. This is understandable since membrane proteins are notoriously difficult to characterize experimentally [16], which limits their number in the Protein Data Bank (PDB) [17] and also decreases the probability of obtaining both bound and unbound conformations of the structures that make up the complex. The latter is one of the requirements of any docking benchmark to allow a realistic evaluation of docking performance. We have however reached a point where enough structures of membrane proteins have been deposited in the PDB to create a docking benchmark. This allows us here to introduce a new membrane protein–protein complex benchmark, establish the baseline performance of HADDOCK in two docking scenarios and provide a decoy data set that will allow further optimization of scoring functions for this specific class of complexes. This new benchmark is freely available for download. The structures have been renumbered and cleaned to facilitate immediate docking and analysis. In addition to the benchmark itself, we are also providing code that can be used for the analysis of docking results as well as the decoy data sets. The content of this benchmark is more diverse than any of the data sets used in previous studies since it contains both non-helical proteins and helical proteins, and complexes that are larger than GPCRs, as well as small helices.

Materials and Methods

Data sources

The primary data source for this benchmark was the MPSTRUC database (http://blanco.biomol.uci.edu/mpstruc/). MPSTRUC is a manually curated database of membrane proteins. Its entries are classified into three categories:

1. Monotopic membrane proteins
2. Beta-barrel transmembrane proteins
3. Alpha-helix transmembrane proteins

We disregarded the monotopic membrane protein category since it is made up of proteins, which are not embedded in the lipid bilayer but instead are only anchored to one side of it. We considered all remaining unique entries and processed them using the procedure outlined in Fig. 1.

After identifying a complex, we searched the related structures in MPSTRUC as well as the homologous structures of that complex in the PDB to identify potential unbound structures of its components. The related MPSTRUC entries correspond to the same protein structure solved under different conditions (e.g., acidic *versus* basic pH), with different techniques (NMR *versus* X-RAY crystallography) or complexed with other biomolecules (e.g., small-molecule ligands or peptides). For the complexes where we could not identify a suitable unbound structure via MPSTRUC, we turned to the PDB and made use of its precalculated sequence similarity clustering analysis results. Optimally, the structure of the complex and that of its components should have been determined independently of each other and be complete, that is, have no missing parts or mutations close to the interface. If that is not the case, but highly homologous structures are available, those are included instead. In this case, highly homologous refers to 100% sequence identity (without gaps) of the interface region and very similar (if not identical) sequence for the remainder of the protein. In these cases, the remainder of the protein was not modeled since the overall similarity is quite high. SI Table 1 lists the backbone RMSD (after optimal superimposition using backbone atoms) of all components for all entries that are not classified as "bound" (see Table 1). The mean RMSD is 1.45 ± 0.86 Å. If the homologous structures differ at the interface—due to

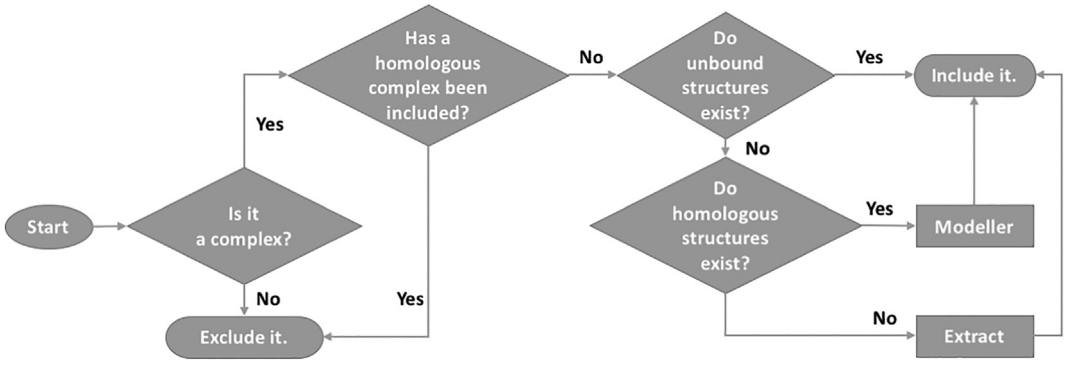

Fig. 1. Flowchart of the structure identification procedure.

Table 1. The dimeric entries of the membrane protein complex docking benchmark

Complex	Unbound PDB ID 1	Unbound PDB ID 2	Category	Composition	Difficulty	i-RMSD (Å)	Buried surface area (Å^2)	Secondary structure
2bg9	2bg9_ADE	2bg9_BC	Both	BB	Bound	0	5452.5	Helical
2bs2	2bs2_AB	2bs2_CD	MS	BB		0	4173.9	Helical
2r6g-TM	2r6g_F	2r6g_G	TM	BB		0	8073.3	Helical
2vpz	2vpz_AB	2vpz_CD	MS	BB		0	2064.7	Helical
4hg6	4hg6_A	4hg6_B	TM	BB		0	4704.2	Helical
4huq-TM	4huq_S	4huq_T	TM	BB		0	5202.9	Helical
4huq-TM-A	4huq_ST	4huq_A	MS	BB		0	1771.8	Helical
4huq-TM-B	4huq_ST	4huq_B	MS	BB		0	2682.6	Helical
5a63-BC	5a63_B	5a63_C	TM	BB		0	3430.2	Helical
2hdi	2hdi_A	1cii_A	Buried	UB	Easy	0.361	1925.7	Beta
4j3o	4j3o_D	3bfq_FG	Buried	UB		0.392	4681.2	Beta
1 m56	2gsm_AB	1 m56_CD	TM	UB		0.572	4961.5	Helical
1k4c	1k4c_A	1j95_ABCD	MS	UU		0.638	1766.9	Helical
3x29	3x29_A	2quo_A	MS	UB		0.673	2143.3	Helical
2k9j	2rmz_A	2k1a_A	TM	UU		0.678	982.0	Helical
2r6g-TM-peri	2r6g_FG	1jw4_A	MS	UB		0.716	3807.0	Helical
2gsk	2guf_A	1u07_A	MS	UU		0.86	1636.2	Beta
5aww	5aww_YG	5aww_E	TM	UB		0.868	2636.5	Helical
2zxe-AG	2zxe_A	2zxe_G	TM	UB		0.919	1528.0	Helical
2zxe-AB	2zxe_A	2zxe_B	TM	UB		0.94	1503.5	Helical
3wxw	3wxw_A	1vfa_HL	AB	HB		0.982	1672.9	Helical
3hd7	3hd7_A	3hd7_B	TM	UU	Intermediate	1.024	663.2	Helical
3csl	3csl_A	1b2v_A	MS	UB		1.065	3681.6	Beta
2ks1	2n2a_A	2m0b_A	TM	UU		1.158	662.2	Helical
5d0o	5d0o_A	2yhc_A	MS	UB		1.182	2909.4	Beta
5a63-AC	5a63_A	5a63_C	TM	UB		1.218	1953.1	Helical
3p0g	2rh1_A	4unu_A	AB	UU		1.26	1801.8	Helical
2r6g-TM-cyto	2r6g_FG	1q12_AB	MS	UB		1.363	3959.9	Helical
3o0r	3o0r_B	3o0r_C	AB	UB		1.445	2383.6	Helical
5fxb	5fxb_AB	1ttf_A	AB	HB		1.475	1716.6	Helical
4q35	4q35_A	4nhr_A	Buried	UB	Hard	2.061	5592.7	Beta
4 m48	4 m48_A	4dvb_HL	AB	HB		2.335	1144.1	Helical
2hi7	1ti1_A	2k73_A	MS	UU		2.588	1337.9	Helical
1ots	1kpk_AB	4nzu_HL	AB	HU		3	1327.7	Helical
3v8x	3v8x_A	4x1b_A	MS	HB		3.422	4945.0	Beta

The first column is the PDB ID of the complex structure, and columns 2 and 3 are the PDB IDs of the unbound structures; category refers to the complex type; composition refers to the origin of every component of the complex; difficulty and I-RMSD reflect the difficulty of the target; secondary structure classifies the complex into one of two categories (beta and helical) depending on the secondary structure characteristics of its transmembrane domain; and buried surface area refers to the buried surface area at the interface of every complex. The categories are buried, MS, TM, both and AB, and they correspond to complexes whose interface lies inside a β-barrel, between cytosolic and transmembrane domains, between transmembrane domains, between transmembrane–cytosolic and transmembrane–transmembrane domains, and a complex of an antibody-like domain that stabilizes a transmembrane domain, respectively. The composition types can be BB, UB, HB, HU and UU, and they stand for bound–bound, unbound–bound, homology–bound, homology–unbound, and unbound–unbound, respectively. BB means that both chains originate in the bound complex, UB means that one of the chains originates in the bound complex and the other in another structure, HB means that one chain is a homology model based on another structure/complex and the other originates from the bound complex, HU means that one chain is a homology model based on another structure/complex and the other originates from another complex or free structure, and UU means that both chains originate from another structure.

mutations or gaps—they were modeled with modeller [18], using the *loopmodel* protocol for the cases with significant interface gaps and the *automodel* protocol for all remaining ones. We made use of homology models for five complexes (see Table 1) for which the sequence similarity and identity ranged between 71% and 96% and between 56% and 96%, respectively. One hundred models were generated and ranked according to their objective function score, and the best-scoring structure that was within modeling difficulty of the complex structure was selected after visual inspection. We manually inspected the models to ensure no unnatural segments were introduced during the modeling of the gaps.

We applied additional selection criteria: we only selected heteromeric interfaces; therefore, homomeric complexes that function as multi-chain proteins such as trimeric transmembrane porins (e.g., PDB entry 1OSM [19]), although technically transmembrane protein complexes, were not included. X-ray structures were given priority over structures determined by NMR, and higher-quality structures (resolution, clash-score, R-free, Ramachandran outliers) were preferred over lower-quality structures. The availability of high-quality unbound structures also influenced the inclusion of one complex over another for which no unbound structures were available or, if there were, they were of low quality (low resolution, mutations, gaps). The

resulting data set is also non-redundant in the sense that we have only included what we determined as the best complex based on the above criteria for any given protein family. In addition to the MPSTRUC classifications, we also made use of a sequence identity cutoff of 30% for identifying homologous structures to ensure we only included non-redundant entries. Accordingly, no chain of any complex of the data set has a sequence identity larger than 30% to any other. For calculating the sequence identities, we used the Needleman–Wunsch algorithm [20] with the BLOSUM62 [21] matrix, and a gap open and extend penalty of 10 and 0.5, respectively.

The entries of the benchmark have been modified to facilitate comparisons between the unbound and reference structures. The numbering and chains ids of the unbound structures have been modified to match those of the reference structures. Disordered regions were removed when near the interface or when they introduced challenging conformational rearrangements that would prevent the unbound structures from adopting a conformation close to the reference one. We have made our best efforts to include all biologically relevant ions and cofactors when they were present in both unbound and reference structures. In some cases, we have joined two or more unbound chains in a single body. Reasons for doing so are reducing the number of docking partners to two or three, since most docking codes do not support multi-body docking, the availability of unbound structures, and the topology of the complex—an example would be joining two homomeric TM subunits in a single subunit and docking that against a cytosolic partner. These cases are indicated by the presence of multiple chain ids at the end of the unbound structure id in Table 1. We consider different subunits of the complex for the four complexes (2r6g, 2zxe, 4huq, 5a63), which appear more than once (see Table 1). We only used the renamed and renumbered unbound and homology structures for docking in all cases where such structures were available. In all other cases, we used the renamed and renumbered bound structures.

Docking

The HADDOCK webserver (v2.2) (https://haddock.science.uu.nl/services/HADDOCK2.2) [22] was used for all docking runs. HADDOCK is an integrative modeling biomolecular docking platform that makes use of experimental data (mostly derived from biophysical/biochemical experiments) or bioinformatics predictions to drive the docking process. This information is typically translated into distance restraints used to drive the docking [23]. The docking consists of three stages:

i. Rigid-body energy minimization—it0
ii. Semi-flexible refinement by simulated annealing in torsional space—it1
iii. Refinement in explicit solvent—itw

For the first stage (it0), the partners are randomly oriented and translated away from each other followed by rigid-body energy minimization. For it1, flexibility is introduced in the interface residues of the complex (defined as the set of residues whose atoms are within 5 Å of any atom of any partner), first along the side-chains only and, in the final stage, including the backbone atoms as well. The last stage (itw) consists of a short molecular dynamics run in Cartesian space and explicit solvent (the docking runs were performed with the default TIP3P water model [24]).

We used two types of restraints to drive the docking: random and true interface restraints. In the case of random restraints, for each docking trial, a surface-exposed patch of residues is randomly defined on both partners of a dimeric complex and used to drive the docking by defining those patches as active residues in the HADDOCK formalism. Since this option is not supported for higher-order complexes, for the three trimeric complexes in the benchmark (see Results and Discussion) center-of-mass, C3 symmetry and non-crystallographic symmetry restraints were used instead [25]. In the case of true interface restraints, we extracted the interface residues of the bound complex (at a distance cutoff of 5 Å) and defined those as active in HADDOCK for the docking run.

The number of docking decoys generated was set for it0/it1/itw to 50,000/400/400 and 10,000/400/400, for *ab initio* (random restraints) and true interface-driven docking, respectively. In addition, since the scoring function of HADDOCK has not been optimized yet for membrane complexes, we set the number of trials in it0 to 1 and disabled the systematic sampling of 180° rotations during it0 to 1 to disable the internal scoring scheme of HADDOCK. For the cases categorized as "buried" (see Table 1), we have also lowered the intermolecular energy scaling to 0.01 to allow interpenetration of chains during it0. We further kept the original scoring function of HADDOCK, defined as:

$$\text{HS–it0} = 0.01 * E_{\text{vdw}} + 1.0 * E_{\text{elec}} + 1.0 * E_{\text{desolv}} + 0.01 * E_{\text{AIR}} - 0.01 * \text{BSA}$$

$$\text{HS–it1} = 1.0 * E_{\text{vdw}} + 1.0 * E_{\text{elec}} + 1.0 * E_{\text{desolv}} + 0.1 * E_{\text{AIR}} - 0.01 * \text{BSA}$$

$$\text{HS–itw} = 1.0 * E_{\text{vdw}} + 0.2 * E_{\text{elec}} + 1.0 * E_{\text{desolv}} + 0.1 * E_{\text{AIR}}$$

where E_{vdw}, E_{elec}, E_{desolv} and E_{AIR} stand for van der Waals, electrostatic, desolvation and restraint energies. The non-bonded terms are calculated with the OPLS force field [26] with a cutoff of 8.5 Å; the desolvation parameters are described in Ref. [27] and

the restraint energy in Ref. [23]. BSA stands for buried surface area. It is worth noting that the desolvation potential depends on parameters that have been optimized for soluble proteins. These are thus the default scoring settings for soluble complexes.

Analysis

We report both the interface and ligand RMSD values (I/L-RMSD, respectively) as used in CAPRI. For the I-RMSD, we superimpose and calculate the RMSD of the backbone atoms of the interface residues (defined at a 10-Å cutoff). For the L-RMSD, we superimpose on the backbone atoms of the receptor (defined as the largest of the partners) and calculate the RMSD of the backbone atoms of the ligand (defined as the smallest of the partners). For the 2 trimers (see Table 2), I-RMSD is calculated as described above and L-RMSD is calculated by selecting the first chain as the receptor and averaging the L-RMSD of the second and third chains. Fitting and RMSD calculations were performed using the McLachlan algorithm [28] as implemented in the program ProFit (http://www.bioinf.org.uk/software/profit/) from the SBGrid distribution [29]. All scripts used for analysis are provided together with the docking benchmark at https://github.com/haddocking/MemCplxDB.

Results and Discussion

Benchmark

Following the protocol that is outlined in the Materials and Methods section, we identified 37 complexes of interest. These complexes are listed in Tables 1 and 2 (dimers and trimers, respectively). An annotated version of this table, detailing the modifications that were made to the structures, can be found in the SI (SI Tables 2 and 3). The tables detail the PDB ID of the structure of the complex and those of the corresponding unbound entries. In the cases where all the components have been extracted from the complex, that entry is a defined as a "bound" case. If at least one of the partners is not extracted from the complex, then that case is classified as "unbound" and, depending on the I-RMSD of the unbound structures after optimal superposition on the reference, is classified as "easy," "intermediate" or "hard," difficulty based on I-RMSD values of less than 1 Å, between 1 and 2 Å and over 2 Å, respectively. Both trimeric entries of the benchmark (Table 2) are classified as "unbound" because the "unbound" components originate from a different PDB entry of the same complex crystalized under different conditions (3w9h and 2qts for 2j8s and 4fz0, respectively). Those differ significantly enough in their i-RMSD (0.65 and 1.18 Å for 2j8s and 4fz0, respectively) and overall backbone RMSDs of each subunit (see Table S1) from what we define as the reference bound conformation, which justifies their inclusion in this benchmark. We have also categorized the complexes based on the nature of the interaction. Complexes whose interface is contained within the membrane are labeled "TM" for transmembrane, and complexes whose interface lies between the membrane and the cytosolic/periplasmic/extracellular environment are labeled "MS" for membrane-soluble. Complexes whose interface lies in the membrane but also extends past it are labeled "both" for both transmembrane and membrane-soluble. Complexes where one of the partners is embedded in a TM beta-barrel are labeled "buried" and are by nature TM complexes, and complexes that involve antibodies, antibody fragments, monobodies or nanobodies are labeled "AB" and are by nature MS complexes. For details regarding the benchmark assembly, refer to the Materials and Methods section and SI Tables 2 and 3.

Docking

To establish the baseline performance of HADDOCK on this membrane protein complex docking benchmark and generate a docking decoy data set that can serve for further optimization of membrane-specific

Table 2. The trimeric entries of the membrane protein complex docking benchmark

Complex	Unbound PDB ID 1	Unbound PDB ID 2	Unbound PDB ID 3	Category	Composition	Difficulty	i-RMSD (Å)	Buried surface area (Å^2)	Secondary structure
2j8s	3w9h_A	3w9h_B	3w9h_C	Both	UUU	Easy	0.648	10,358.8	Helical
4fz0	2qts_A	2qts_B	2qts_C	Both	UUU	Intermediate	1.18	12,084.0	Helical

The first column is the PDB ID of the complex structure, and columns 2, 3 and 4 are the PDB IDs of the unbound structures; category refers to the complex type (refer to Table 1 for details); composition refers to the origin of every component of the complex (refer to Table 1 for details); difficulty and I-RMSD reflect the difficulty of the target; secondary structure classifies the complex into one of two categories (beta and helical) depending on the secondary structure characteristics of its transmembrane domain; and buried surface area refers to the buried surface area at the interface of every complex. The unbound subunits for both entries originate from a different PDB entry of the same complex crystalized under different conditions (3w9h and 2qts for 2j8s and 4fz0, respectively). Both their individual backbone RMSDs (see Table S1) and i-RMSD values justify their classification as "UUU".

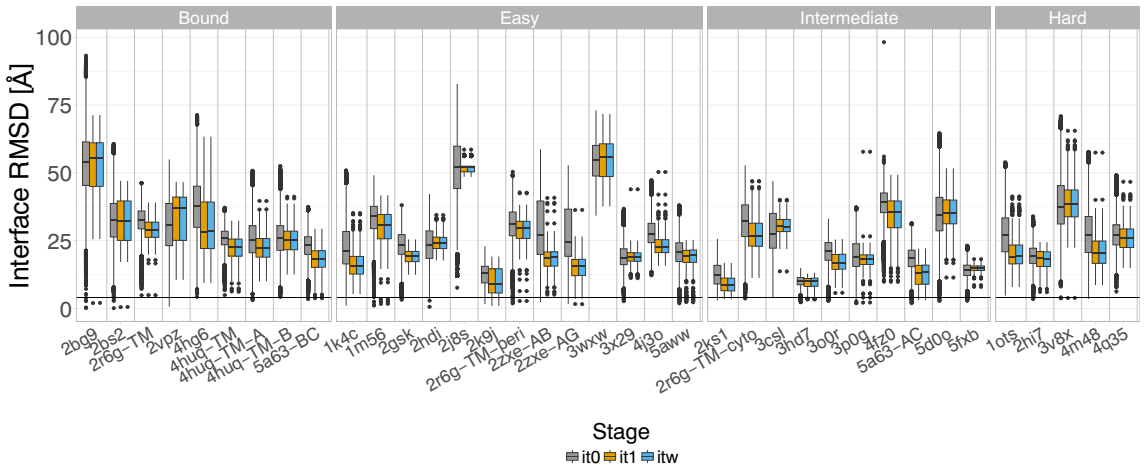

Fig. 2. I-RMSD values of the docking decoys of the membrane protein complex docking benchmark for the random-restraint driven runs. The complexes are grouped by difficulty. Each complex is represented by three boxplots, corresponding to the rigid-body (gray), semi-flexible refinement (orange) and final water refinement (blue) stages of HADDOCK. The black line represents the acceptability cutoff of 4 Å I-RMSD. The boxes of the boxplots range from the first to the third quartile, the upper whisker extends from the hinge to the maximum value or 1.5 * Interquantile range (IQR), the lower whisker extends from the hinge to the minimum value or 1.5 * IQR, and outliers are shown as black points.

scoring functions, we performed the docking using two different scenarios.

Random restraints

In the first scenario, HADDOCK was used in its *ab initio* mode with random restraints. Figure 2 shows the distribution of I-RMSD values for all three stages of the docking runs (SI Fig. 1 shows the same plot but for L-RMSD). The RMSD values have been calculated according to CAPRI criteria as specified in the Materials and Methods section. The boxplots colored gray, orange and cyan correspond to the I-RMSD values of it0, it1 and itw, respectively. The horizontal black line

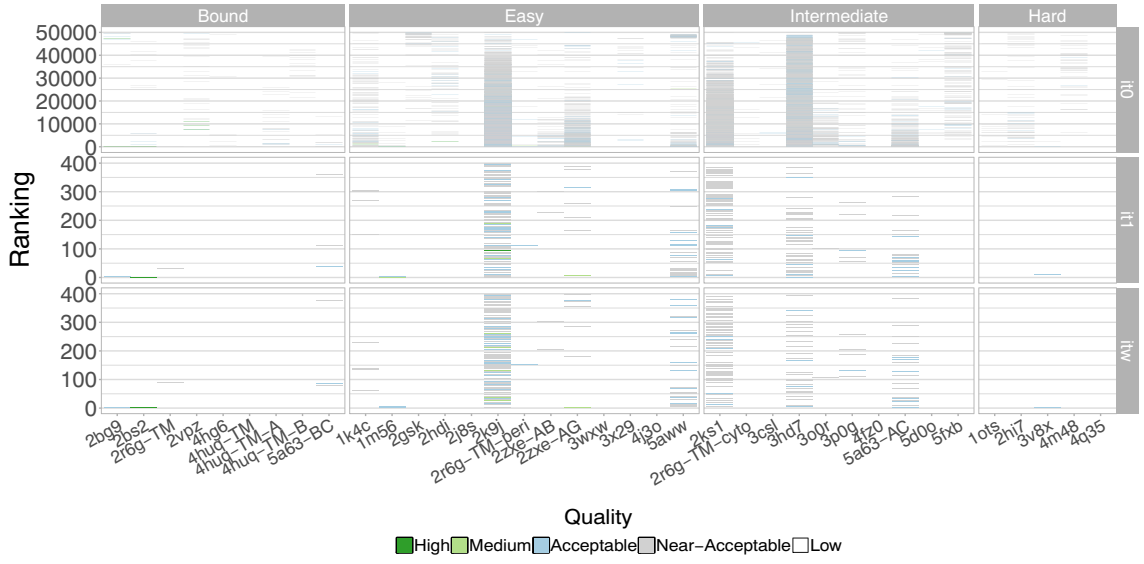

Fig. 3. Quality assessment of the generated models of the random-restraint driven runs based on I-RMSD values. The complexes are grouped by difficulty. For each, results of the rigid-body docking (it0; top panel), semi-flexible refinement (it1; middle panel) and water refinement (itw; bottom panel) are shown. The *Y* axis for all sub-graphs corresponds to the ranking of the models according to the default HADDOCK scoring function, with models ranked near 0 having the best scores. Every model has been colored according to its quality, with high-, medium-, acceptable, near-acceptable and low-quality models having I-RMSD values of less than 1 Å (dark green), between 1 and 2 Å (light green), between 2 and 4 Å (light blue), between 4 and 6 Å (light gray) and over 6 Å (white), respectively.

represents the acceptability cutoff of 4 Å. Figure 3 shows the same information, but instead of displaying the raw I-RMSD values, the models are classified by their quality. The figure is separated into 12 sub-graphs, each of which corresponds to one of the three docking stages (it0, it1 and itw) in one of the four difficulty groups (bound, easy, intermediate and hard). Every sub-graph groups the performance for all complexes that have been classified into the same difficulty category for one of the three docking stages. The *Y* axis corresponds to the ranking of every model according to its HADDOCK score as calculated by the appropriate scoring function

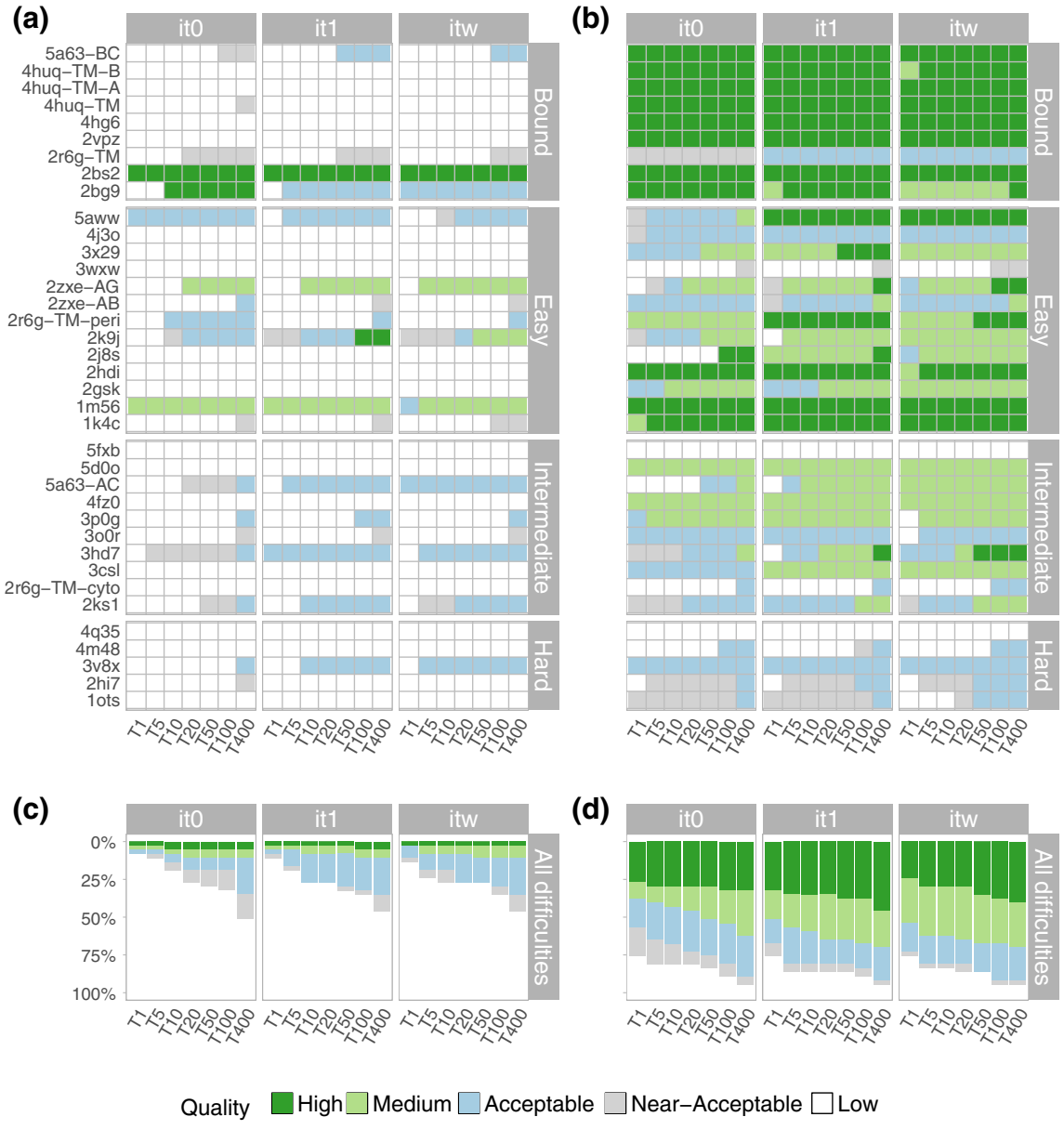

Fig. 4. Evaluation of the success rate as a function of the number of models considered. Every set of horizontal cells corresponds to the performance of HADDOCK on a given complex, with the performance for random restraints being shown in panels a and c, and the performance for true interface restraints in panels b and d. For the top panels, every cell corresponds to the quality of the best model (in terms of I-RMSD) when considering the *N* best models (*N* having the values 1, 5, 10, 20, 50, 100 and 400) for all three docking stages, with the coloring of the cell representing high-, medium-, acceptable and near-acceptable quality models [I-RMSD values of less than 1 Å (dark green), between 1 and 2 Å (light green), between 2 and 4 Å (light blue), between 4 and 6 Å (light gray) and over 6 Å (white), respectively]. Cells that correspond to cutoffs where only low-quality models were generated are colored white. In the bottom panels, the success rate percentage for all complexes is shown as a function of the number of models considered. The coloring is the same as for the top panel.

for every stage (see Materials and Methods), with models ranked near the bottom having a better score. Every model is represented by a single horizontal bar, with the color of the bar representing the quality of the model. There are a limited number of acceptable models since we only used random restraints (*ab initio* docking mode) to drive the docking. Despite that, HADDOCK was able to generate at least one acceptable model in 27 of 37 (~73%) cases during it0 when considering all 50,000 generated models. In 13 of those cases, at least one acceptable model was also selected in the top 400, which are selected for further refinement in it1 and itw. This means that our scoring function could identify at least one acceptable model in ~48% of the cases where at least one model of acceptable quality was generated during it0. The success rate of 48% might sound less than ideal, but it becomes more impressive when one considers the number of acceptable models generated against the size of the sampling pool: In most cases, only a few ($\leq$10) acceptable models were generated in it0 and they were correctly identified as near-native among 50,000 models. SI Table 4 lists the number of acceptable structures generated during it0 for all complexes as well as the number of acceptable complexes ranked in the top 400. No more than these few acceptable models were sampled for the majority of complexes; however, near-native structures were identified even when there were less than 5 of those in a pool of 50,000 (1m56, 2bs2), including one case where the single near-native complex generated was selected (3v8x). The overall success rate of HADDOCK at the water stage using random restraints is ~35%, when considering all water models, as models of acceptable quality were generated in 13 of 37 cases. The difficulty or category of a complex seems to have no effect on the performance of HADDOCK with all difficulties and categories proportionately represented in the list of complexes for which no acceptable model was generated. The distribution of success rates when considering different cutoffs, for the random-restraint driven runs, can be seen in the left panel of Fig. 4. Every cell of that plot corresponds to the quality of the best model (minimum I-RMSD) when considering the top *N* structures (with *N* being 1, 5, 10, 20, 50, 100 and 400). The color of the plot represents the quality of the minimum I-RMSD model. Although the number of complexes for which at least one acceptable model was generated in the top400 did not change between the rigid body and refinement stages, refinement did improve the ranking of those acceptable models as well as their quality. This trend is also reflected in the mean rank of the first acceptable model, which is ~110, ~22 and ~33 for it0, it1 and itw, respectively. The mean I-RMSD of all acceptable models is 3.38 ± 0.5, 3.15 ± 0.77 and 3.13 ± 0.76 Å for it0, it1 and itw, respectively.

True interface restraints

Figure 5 shows the performance of HADDOCK when using true interface information from the native complex to drive the docking, which thus represents an ideal scenario (SI Fig. 2 shows the same plot but for L-RMSD). Unlike for the random restrains runs, the difficulty of the complex is now important and is the main limiting factor for the performance of these runs. This is particularly apparent when comparing the bound and hard targets. In the bound complexes, both sampling and scoring are better since a greater number of high-quality structures are generated during it0 and are scored near the top, meaning

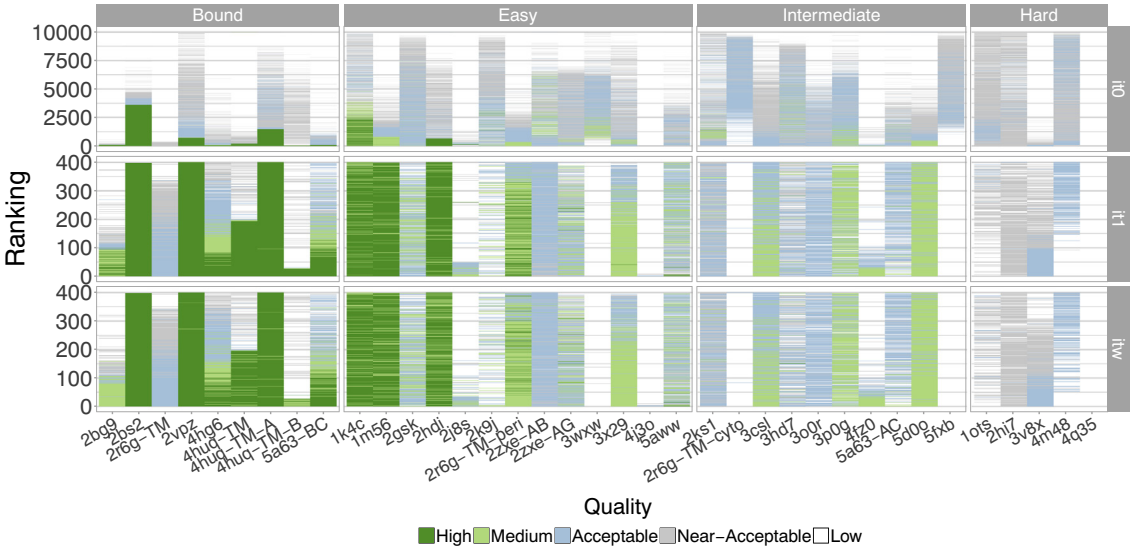

Fig. 5. Quality assessment of the generated models of the true-interface restraint driven runs based on I-RMSD values. For details refer to the caption of Fig. 3.

they proceed to the refinement stages. In general, the performance of HADDOCK is excellent, with 36 of 37 complexes having at least one acceptable (I-RMSD $\leq$ 4 Å) or near-acceptable (I-RMSD $\leq$ 6 Å) model in the rigid-body stage. The inclusion of near-acceptable models can be justified by the fact that when using a well-defined set of restraints, a rigid-body model in the near-acceptable range might become acceptable after semi-flexible refinement, as is the case, for example, for complex 2r6g-TM. For that complex, no acceptable models were generated during it0, but the scoring function successfully identified the best models and, after refinement, more than half of the it1 and itw models became acceptable. Acceptable models are generated for 34 of 37 complexes during the refinement stages corresponding to an overall success rate of 92%, when considering all water models. The right panel of Fig. 4 shows the distribution of success rates for different cutoffs (see "random restraints" above for more details). Except for three cases (2r6g-TM-cyto, 3wxw and 5fxb), for which acceptable models were generated in it0 but not scored in the top 400 that are selected for semi-flexible refinement, our scoring function works well, ranking most near-native models higher than the non-native ones.

Benchmark availability

The bound and unbound structures, including the renumbered models used for docking of the membrane protein complex docking benchmark version1, along with ProFit analysis scripts can be freely downloaded at https://github.com/haddocking/MemCplxDB. The HADDOCK docking decoys are made available through the SBGrid Data Bank [30] and can be downloaded at https://data.sbgrid.com/618 [31].

Conclusion

We have assembled a membrane protein–protein docking benchmark that, to the best of our knowledge, is the first of its kind. The benchmark is freely available for download from GitHub and, in addition to the reference and unbound structures, includes renumbered, docking-ready structures, reference structures and analysis scripts for the calculation of the RMSD metrics that we are reporting in this paper. We have established the docking performance baseline of HADDOCK for two extreme scenarios. Despite the fact that HADDOCK has not been optimized for membrane proteins, it demonstrates excellent performance in the case where high-quality interface data are available, with a 92% overall success rate when considering all 400 itw models. In its *ab inito* docking mode, however, the performance drops to 35% for itw models. In particular, the sampling performance in the rigid body docking stage is affected, where we generate at least one acceptable model in 73% of the cases but only select at least one for further refinement in 48% of them, with many near native models not being selected for the semi-flexible refinement stage as a result. This leaves room for optimization. All docking decoys for the various stages and scenarios can be freely downloaded from the SBGrid data bank. This new docking benchmark and its associated docking decoys should be a valuable resource for the community to foster the development of docking and scoring approaches for membrane protein complexes.

Acknowledgments

This work was supported by the European H2020 e-Infrastructure grants West-Life (grant no. 675858) and BioExcel (grant no. 675728) and by the Dutch Foundation for Scientific Research (NWO) (TOP-PUNT grant 718.015.001). The authors thank Dr. Irina Moreira for helpful discussions.

Conflict of interest statement

The authors declare no competing financial interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.11.005.

Received 19 July 2018;
Received in revised form 2 November 2018;
Accepted 5 November 2018
Available online 9 November 2018

Keywords:
docking;
membrane proteins;
protein–protein complexes;
scoring;
HADDOCK

†I.F. and C.W.v.N. contributed equally to this work.

Abbreviations used:
PDB, Protein Data Bank; MPSTRUC, Membrane Proteins of Known 3D Structure.

References

[1] J. Janin, K. Henrick, J. Moult, L. Ten Eyck, M.J.E. Sternberg, S. Vajda, I. Vakser, S.J. Wodak, CAPRI: a critical assessment of PRedicted interactions, Proteins 52 (2003) 2–9, https://doi.org/10.1002/prot.10381.

[2] S. Gathiaka, S. Liu, M. Chiu, H. Yang, J.A. Stuckey, Y.N. Kang, J. Delproposto, G. Kubish, J.B. Dunbar, H.A. Carlson, S.K. Burley, W.P. Walters, R.E. Amaro, V.A. Feher, M.K. Gilson, D3R grand challenge 2015: evaluation of protein—ligand pose and affinity predictions, J. Comput. Aided Mol. Des. 30 (2016) 651–668, https://doi.org/10.1007/s10822-016-9946-8.

[3] Z. Gaieb, S. Liu, S. Gathiaka, M. Chiu, H. Yang, C. Shao, V.A. Feher, W.P. Walters, B. Kuhn, M.G. Rudolph, S.K. Burley, M. K. Gilson, R.E. Amaro, D3R grand challenge 2: blind prediction of protein—ligand poses, affinity rankings, and relative binding free energies, J. Comput. Aided Mol. Des. 32 (2018) 1–20, https://doi.org/10.1007/s10822-017-0088-4.

[4] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, Z. Weng, Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2, J. Mol. Biol. 427 (2015) 3031–3041, https://doi.org/10.1016/j.jmb.2015.07.016.

[5] M. Trellet, A.S.J. Melquiond, A.M.J.J. Bonvin, A unified conformational selection and induced fit approach to protein–peptide docking, PLoS One 8 (2013) https://doi.org/10.1371/journal.pone.0058769.

[6] N. London, D. Movshovitz-Attias, O. Schueler-Furman, The structural basis of peptide-protein binding strategies, Structure 18 (2010) 188–199, https://doi.org/10.1016/j.str.2009.11.012.

[7] M. van Dijk, A.M.J.J. Bonvin, A protein–DNA docking benchmark, Nucleic Acids Res. 36 (2008) https://doi.org/10.1093/nar/gkn386.

[8] M.J. Hartshorn, M.L. Verdonk, G. Chessari, S.C. Brewerton, W.T.M. Mooij, P.N. Mortenson, C.W. Murray, Diverse, high-quality test set for the validation of protein–ligand docking performance, J. Med. Chem. 50 (2007) 726–741, https://doi.org/10.1021/jm061277y.

[9] R.F. Alford, J. Koehler Leman, B.D. Weitzner, A.M. Duran, D.C. Tilley, A. Elazar, J.J. Gray, An integrated framework advancing membrane protein modeling and design, PLoS Comput. Biol. (2015) https://doi.org/10.1371/journal.pcbi.1004398.

[10] S. Chaudhury, M. Berrondo, B.D. Weitzner, P. Muthu, H. Bergman, J.J. Gray, Benchmarking and analysis of protein docking performance in Rosetta v3.2, PLoS One (2011) https://doi.org/10.1371/journal.pone.0022477.

[11] N. Hurwitz, Di. Schneidman-Duhovny, H.J. Wolfson, Memdock: an α-helical membrane protein docking algorithm, Bioinformatics (2016)https://doi.org/10.1093/bioinformatics/btw184.

[12] S. Viswanath, D.V.S. Ravikant, R. Elber, DOCK/PIERR: Web server for structure prediction of protein-protein complexes, Methods Mol. Biol. (2014) https://doi.org/10.1007/978-1-4939-0366-5_14.

[13] S. Viswanath, L. Dominguez, L.S. Foster, J.E. Straub, R. Elber, Extension of a protein docking algorithm to membranes and applications to amyloid precursor protein dimerization, Proteins Struct. Funct. Bioinf. (2015) https://doi.org/10.1002/prot.24934.

[14] M.A. Lomize, A.L. Lomize, I.D. Pogozheva, H.I. Mosberg, OPM: orientations of proteins in membranes database, Bioinformatics (2006) https://doi.org/10.1093/bioinformatics/btk023.

[15] D.R. Weiss, A. Bortolato, B. Tehan, J.S. Mason, GPCR-bench: a benchmarking set and practitioners' guide for G protein-coupled receptor docking, J. Chem. Inf. Model. 56 (2016) 642–651, https://doi.org/10.1021/acs.jcim.5b00660.

[16] J.G. Almeida, A.J. Preto, P.I. Koukos, A.M.J.J. Bonvin, I.S. Moreira, Membrane proteins structures: a review on computational modeling tools, Biochim. Biophys. Acta Biomembr. 1859 (2017) 2021–2039, https://doi.org/10.1016/j.bbamem.2017.07.008.

[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[18] A. Šali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, J. Mol. Biol. 234 (1993) 779–815, https://doi.org/10.1006/jmbi.1993.1626.

[19] R. Dutzler, G. Rummel, S. Albertí, S. Hernández-Allés, P.S. Phale, J.P. Rosenbusch, V.J. Benedí, T. Schirmer, Crystal structure and functional characterization of OmpK36, the osmoporin of *Klebsiella pneumoniae*, Structure 7 (1999) 425–434, https://doi.org/10.1016/S0969-2126(99)80055-0.

[20] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. (1970) https://doi.org/10.1016/0022-2836(70)90057-4.

[21] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. (1992) https://doi.org/10.1073/pnas.89.22.10915.

[22] G.C.P. van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries, A.M.J.J. Bonvin, The HADDOCK2.2 Web server: user-friendly integrative modeling of biomolecular complexes, J. Mol. Biol. 428 (2016) 720–725, https://doi.org/10.1016/j.jmb.2015.09.014.

[23] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, HADDOCK: a protein–protein docking approach based on biochemical or biophysical information, J. Am. Chem. Soc. 125 (2003) 1731–1737, https://doi.org/10.1021/ja026939x.

[24] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926, https://doi.org/10.1063/1.445869.

[25] E. Karaca, A.S.J. Melquiond, S.J. de Vries, P.L. Kastritis, A. M.J.J. Bonvin, Building macromolecular assemblies by information-driven docking, Mol. Cell. Proteomics 9 (2010) 1784–1794, https://doi.org/10.1074/mcp.M000051-MCP201.

[26] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, J. Am. Chem. Soc. 110 (1988) 1657–1666, https://doi.org/10.1021/ja00214a001.

[27] J. Fernandez-Recio, M. Totrov, R. Abagyan, Identification of protein-protein interaction sites from docking energy landscapes, J. Mol. Biol. 335 (2004) 843–865.

[28] A.D. McLachlan, Rapid comparison of protein structures, Acta Crystallogr. Sect. A Cryst. Phys. Diffr. Theor. Gen. Crystallogr. 38 (1982) 871–873 http://scripts.iucr.org/cgi-bin/paper?S0567739482001806.

[29] A. Morin, B. Eisenbraun, J. Key, P.C. Sanschagrin, M.A. Timony, M. Ottaviano, P. Sliz, Collaboration gets the most out of software, elife 2013 (2013) https://doi.org/10.7554/eLife.01456.

[30] P.A. Meyer, S. Socias, J. Key, E. Ransey, E.C. Tjon, A. Buschiazzo, M. Lei, C. Botka, J. Withrow, D. Neau, K. Rajashankar, K.S. Anderson, R.H. Baxter, S.C. Blacklow, T.J. Boggon, A.M.J.J. Bonvin, D. Borek, T.J. Brett, A. Caflisch, C.I. Chang, W.J. Chazin, K.D. Corbett, M.S. Cosgrove, S. Crosson, S. Dhe-Paganon, E. Di Cera, C.L. Drennan, M.J. Eck, B.F.

Eichman, Q.R. Fan, A.R. Ferré-D'Amaré, J.C. Fromme, K.C. Garcia, R. Gaudet, P. Gong, S.C. Harrison, E.E. Heldwein, Z. Jia, R.J. Keenan, A.C. Kruse, M. Kvansakul, J.S. McLellan, Y. Modis, Y. Nam, Z. Otwinowski, E.F. Pai, P.J.B. Pereira, C. Petosa, C.S. Raman, T.A. Rapoport, A. Roll-Mecak, M.K. Rosen, G. Rudenko, J. Schlessinger, T.U. Schwartz, Y. Shamoo, H. Sondermann, Y.J. Tao, N.H. Tolia, O.V. Tsodikov, K.D. Westover, H. Wu, I. Foster, J.S. Fraser, F.R. N.C. Maia, T. Gonen, T. Kirchhausen, K. Diederichs, M. Crosas, P. Sliz, Data publication with the structural biology data grid supports live analysis, Nat. Commun. 7 (2016) https://doi.org/10.1038/ncomms10882.

[31] P. Koukos, A. Bonvin, HADDOCK membrane protein–protein complex models, 2018 https://doi.org/10.15785/SBGRID/618.