HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

Proteins. 2016 September ; 84(Suppl 1): 370–391. doi:10.1002/prot.24997.

Biological Function Derived from Predicted Structures in CASP11

Peter J. Huwe[1], **Qifang Xu**[1], **Maxim V. Shapovalov**, **Vivek Modi**, **Mark D. Andrake**, and **Roland L. Dunbrack Jr.**[*]

Fox Chase Cancer Center, Philadelphia PA 19111, USA

Abstract

In CASP11, the organizers sought to bring the biological inferences from predicted structures to the fore. To accomplish this, we assessed the models for their ability to perform quantifiable tasks related to biological function. First, for 10 targets that were probable homodimers, we measured the accuracy of docking the models into homodimers as a function of GDT-TS of the monomers, which produced characteristic L-shaped plots. At low GDT-TS, none of the models could be docked correctly as homodimers. Above GDT-TS of ~60%, some models formed correct homodimers in one of the largest docked clusters, while many other models at the same values of GDT-TS did not. Docking was more successful when many of the templates shared the same homodimer. Second, we docked a ligand from an experimental structure into each of the models of one of the targets. Docking to the models with two different programs produced poor ligand RMSDs with the experimental structure. Measures that evaluated similarity of contacts were reasonable for some of the models, although there was not a significant correlation with model accuracy. Finally, we assessed whether models would be useful in predicting the phenotypes of missense mutations in three human targets by comparing features calculated from the models with those calculated from the experimental structures. The models were successful in reproducing accessible surface areas but there was little correlation of model accuracy with calculation of FoldX evaluation of the change in free energy between the wild type and the mutant.

Keywords

CASP11; protein structure prediction; protein function; protein docking; missense mutation phenotype prediction

Introduction

Protein structure prediction has long been a tool in the biological sciences, since the structures of the first proteins were determined.[1–3] Even in what is widely believed to be the first application of template-based modeling (in 1969), the prediction of the structure of α-lactalbumin from the structure of lysozyme,[4] the model was used to interpret existing experimental data indicating that α-lactalbumin binds (non-catalytically) a shorter

[*]Correspondence: Roland.Dunbrack@fccc.edu.

[1]These authors contributed equally to this work.

polysaccharide than lysozyme due to a truncated binding site. In another early example (1980), Greer used template-based modeling of haptoglobin, a pseudo-protease, based on combining the experimental structures of trypsin, elastase, and chymotrypsin primarily to infer potential binding sites on haptoglobin of an αβ heterodimer of hemoglobin.[5] The model was used subsequently to interpret proteolysis data to identify the hemoglobin binding site,[6] which has only very recently been confirmed with an experimental structure of haptoglobin bound to hemoglobin.[7]

Protein structure prediction, like experimental structure determination, can be used as a key to understanding the biological function of proteins. But the biological function of a protein may have several different meanings. One way to define it for a specific protein is the set of molecules with which the protein interacts and the subsequent biological effects of these interactions. Several examples may be given: a protein may bind DNA at specific sites and enhance transcription through interactions with an RNA polymerase; an enzyme binds its substrate(s) and causes a chemical reaction to occur; a protein may bind to a second protein, and activate (or inhibit) the function of the binding partner, whatever that may be. In some cases, structural information may be helpful in predicting a binding partner or in determining what the downstream effect of binding may be. But it is much more common that structural information is used to determine *how* a protein accomplishes binding to known partners (e.g., what regions of a protein are utilized) and what that binding may do to either or both partners (e.g., conformational or dynamic changes) so that some further effect is accomplished. Such structural information may provide a rationale for the effects of alteration of either protein in the form of post-translational modifications or mutations. Even when structure does not provide a direct answer to these questions, it may be used to generate testable hypotheses and experiments to provide such answers. Given this description of functional inference from structure, in this paper we have utilized data from the CASP11 experiment to evaluate the utility of protein structure prediction in biology and medicine related to binding events and the effects of mutation on binding. Through the Molecular Modeling Facility at the Fox Chase Cancer Center, we have extensive experience in using protein structure prediction in similar tasks in cancer biology.[8–17]

The utility of protein structure prediction for biological inference as described above is difficult to quantify and to benchmark. Many uses of structure prediction do not lend themselves to quantitative assessment, although there are several that do. Yue et al.[18] compared the sequence identity of templates with the accuracy of a structure-based phenotype predictor with template-based modeling with the SCWRL3 program.[19] They found similar accuracies of experimental structures and models for templates with sequence identity above 40%, while accuracy fell especially when sequence identity was below 30%. Yates et al. have recently studied the utility of predicted structures for the prediction of missense mutation phenotypes.[20] Vakser et al. have studied the accuracy of protein-protein docking given predicted structures at different RMSDs from the native monomers[21] and created extensive sets of homology models of monomers at different RMSDs from native suitable for future benchmarking of docking methods.[22] Several groups have studied the accuracy of ligand docking on predicted structures.[23–26] Previous studies have suggested that homology models constructed with >50% sequence identity are accurate enough for structure-based drug design, and those with >30% identity can be used to assess the

druggability of a target.[27,28] Conversely, Bordogna et al. concluded that there is no relationship between model quality and docking accuracy but they suggested that using flexible binding site residues could have a large impact on effectiveness.[24] In most of these studies, the procedure was to utilize a single structure prediction protocol on a set of proteins and to compare the results of functional predictions (docking, phenotypes) with those of the experimental structures. However, few of these studies employed comparison of ensembles of models of the same protein and quantified the relationship of functional predictions with the accuracy of the models.

In previous CASPs, function prediction and ligand-binding site prediction have been assessed. Function prediction was attempted in CASP6[29] and CASP7[30] but turned out to be very difficult given the limited amount of functional information that was available for the CASP targets, many of which were proteins of unknown function from structural genomics projects. In CASP8,[31] CASP9,[32] and CASP10,[33] the experiments were restricted to prediction of binding sites for small ligands, which met with more success since the experimental structures contained ligands so that binding site predictions could be assessed quantitatively.

In CASP11, the organizers sought to emphasize the biological implications of structure prediction by asking the authors of experimental structures in CASP to provide their rationale for pursuing structure determination. In a limited number of targets, this information was provided by the depositors and subsequently given to the predictors who could use the information as they saw fit. The information ranged from interest in the oligomeric structure of the protein, the positions of ligands in the experimental structure, and structural explanations of missense mutations associated with disease. In addition, many targets for which no information was provided by the authors fall into the same categories if publicly available annotations are considered. In this paper, we provide a quantitative analysis between the accuracy of protein structure models and their utility in three common tasks, related to biological function as described above, that can be *quantitatively* evaluated: protein-protein docking; ligand docking; and prediction of missense mutation phenotypes.

We have assessed the utility of models in protein-protein docking by applying the program ClusPro[34,35] on predicted monomer structures to generate models of homodimers, which if successful would be similar to homodimers identified as probable biological assemblies in the crystal structures of CASP11 targets. ClusPro applies fast Fourier transforms to search a six-dimensional space of rotations and translations of the relative positions and orientations of two proteins, and scores docked complexes with a function consisting of attractive and repulsive van der Waals terms, an electrostatic term, and a pairwise amino acid contact term parameterized on known dimer structures.[34] ClusPro is fast, implemented as a web service, and has performed well in blind tests of protein-protein docking.[35] As such, it is highly suitable for the experiment on CASP11 structures that we describe here. To assess the utility of the CASP11 models in docking, we measured the similarity of contacts in the ClusPro homodimer models with those in the experimental structure. The similarity of contacts was measured by the Q score,[36,37] which is a form of Jaccard similarity[38,39] and we investigated the relationship of Q with model accuracy in the form of the GDT-TS score and other CASP accuracy measures.

As a measure of evaluating the usefulness of modeling in drug design, we performed docking calculations on target T0805 with the programs SwissDock[40] and AutoDock Vina,[41] both of which are freely available for academic use as web services. T0805 is an oxygen-insensitive nitroreductase called Rv3368 from *M. tuberculosis*, which has a bound flavodoxin in the unpublished crystal structure. The best templates are 27% identical in sequence. In the docking calculations with AutoDock Vina, we treated both the flavodoxin ligand and the protein receptor as flexible. In the SwissDock calculations, only the ligand was flexible. We compared both RMSD of the docked ligand position from the position in the experimental structure and also the similarity of the contacts of the docked ligand and the native structure based on a Jaccard similarity.

Missense mutation predictors that use protein structure employ a variety of features, but one of the most common is the relative surface accessibility of residues in monomeric or oligomeric structures.[42,43] Since we have lists of disease-associated mutations for three human proteins in CASP11 and few or no neutral mutations in these proteins, we have compared the predicted solvent accessibilities of these mutations with the experimental structures and assessed the root-mean-square deviation of these predictions as a function of structure prediction accuracy scores. As an alternative method of assessing the biophysical consequences of mutations, we calculated the change in Gibbs free energy of mutation using FoldX[44] for each mutation in the predicted structures. FoldX quantitatively estimates the effects of interactions on the stability of proteins and protein complexes. We performed similar calculations on all residues in these three proteins.

Finally, we highlight some predicted structures that generate reasonable hypotheses on how some CASP targets may function in ways that are similar to what one might generate from the experimental structures. In these few cases, the real answer is not yet known, so this analysis is offered only as suggestive.

Results

Homodimer docking of predicted structures of monomers

For two of the CASP11 targets, the predictors were provided with information from the crystallographers that the structure of the dimer was a question posed for the structure determination. In one case, target T0759, the predictors were told that the structure was “Likely monomer, possibly dimer. Related to human cancer.” However, the experimental structure showed that the most likely dimeric interface in the crystal was in fact made up mostly of residues from the His tag, and thus this structure does not provide a suitable target for evaluating docking of models. For target T0792, the predictors were told “Main modeling interest lies in correct prediction of the dimer.” Several other targets are also probable homodimers in their crystal structures, giving us a set of target homodimers to evaluate.

We have focused on using the predictions of the monomer structures to assess the ability of a single docking program to produce good models of the dimers as a function of the accuracy of the structure prediction of the monomer. We excluded several targets that had ambiguous identification of the correct dimer in the crystal and we excluded probable tetramers as well.

This left us with a total of 10 CASP11 template-based-modeling (TBM) targets that turned out to be probable homodimers in the crystal structures, as determined by the authors and/or the program PISA,[45] once they became available for analysis. These are listed in Table I. In all of these cases, the predictors were told that the likely oligomeric state was a dimer. However, all of the targets were available for servers only. Human groups were allowed to submit oligomeric structures of these targets for evaluation by the CAPRI team (Wodak et al., this issue), which we have not analyzed. Thus, we performed docking calculations on server-produced models of the CASP11 targets.

We utilized the ClusPro server[35] to dock two copies of each predicted monomer structure for each target in Table I. ClusPro provide the centroids of approximately 20 of the largest clusters of docked models, sorted in order of decreasing size. We compared these structures with the experimental dimer in the crystal with the Q score that we derived to compare interfaces in multiple crystal forms of homologous proteins.[36,37] Q measures the similarity of contacts in two different protein dimers, whether the proteins are identical in sequence or merely homologues. It penalizes over and under prediction of contacts and is dependent on the position of Cβ atoms (Cα for Gly), not side-chain positions.

The Q score is related to the Jaccard similarity coefficient or Jaccard index, invented by Paul Jaccard in 1901 to measure the overlap of species in different zones of the French alps.[38,39] In his case, the “coefficient of community” was the number of species common to two different zones divided by the number of species in either or both zones in total. The Jaccard index is widely used in bioinformatics and structural biology,[46–48] for instance as a measure of sequence alignment accuracy compared to structure alignments.[49] For comparing two interfaces, Q is the number of contacts that exist in both structures divided by the number of contacts contained in one or both structures. When counting, each contact is scaled by the similarity of the Cβ/Cβ distances in the model and the experimental structure with an exponential function (see Methods). The contacts are also weighted by the shorter of the distance in the predicted structure and the experimental structure, so that closer contacts count more than more distant contacts.

In order to mimic how models might actually be used to predict interfaces in actual applications when the answer is not known, we assessed the top 1, top 5 and top 10 docking models for each CASP predicted structure; we also performed docking with ClusPro on a monomer from each experimental structure. These are reasonable numbers that we would investigate with experimental collaborators, potentially narrowing down the list to 2–3 cases that might be probed by experimental mutagenesis (e.g., selected based on surface area, symmetry, sequence conservation, prior experimental data). The results for the target T0792, for which the dimer was the goal of structure determination,[50] are shown in Figure 1, which demonstrates the relationship between three measures of structure prediction accuracy, GDT-TS (Figure 1A), RMSD (Figure 1B), and LDDT (Figure 1C),[51] and the best Q score of the top 10 cluster centroids provided by ClusPro. The results of using the experimental structure of the T0792 monomer are marked in each panel with a magenta circle. In each figure, we see an L-shaped distribution of points: the poorest models (e.g. GDT-TS<60%) are not able to achieve a good Q score in any of the ClusPro clusters; when GDT-TS is between 60 and 70%, a subset of the models are able to achieve good Q scores (>0.3); however, many of the

models in the same range are not able to provide good predictions of the homodimer structure. Above GDT-TS ~ 73%, four models are all able to achieve high values of Q as does the experimental structure (Q=0.58). The RMSD (Figure 1B) and LDDT plots (Figure 1C) show similar trends, with some structures with RMSD<2.2 Å and/or LDDT>0.5 able to achieve good docking results, while others in the same range do not.

We wondered if the success of docking might be more correlated with the accuracy of the predicted structure of the interface region than it was with the entire structure. We calculated GDT-TS of the interface region (GDT-TS_i) with the program LGA[52] as well as the RMSD of the interface region (RMSD_i). The results (Figures 1D and 1E respectively) are very similar to the GDT-TS and RMSD of the whole structure (Figures 1A and 1B respectively). As expected, GDT-TS and GDT-TS_i are highly correlated (Figure 1F).

The results for the remaining 9 targets are shown in Figures 2, 3, and 4 which show the maximum Q score over the top 10 docked structures for each CASP monomer prediction as a function of GDT-TS (Figure 2) and GDT-TS_i (Figure 3), and the relationship between GDT-TS and GDT-TS_i (Figure 4). The results for the experimental structures are plotted as magenta circles. A common L-shaped pattern is evident in most of the plots in Figures 2 and 3. The poorest models are not able to achieve a good Q score in any of the top clusters, while at some value of GDT-TS (Figure 2) or GDT-TS_i (Figure 3) above 60% (depending on target), many and sometimes most of the predicted monomeric structures are able to do so. However, in all cases, many of the models at the same values of GDT-TS or GDT-TS_i do not produce good homodimer models with ClusPro. The relationship between GDT-TS and GDT-TS_i (Figure 4) shows that for most targets, the interface is more accurate than the entire structure, and for some significantly so (T0852, T0851, T0849). For T0770 and T0843, the GDT-TS of the interface is worse than for the whole structure, and the distribution of Q is notably poor for these targets. The experimental structure (magenta points in Figures 2 and 3) achieved Q scores between 0.3 and 0.7, and in 6 out of 10 cases some models were able to achieve higher values of Q than the experimental structure.

In Figure 5, we show the percent of models that are able to achieve good Q scores in the largest ClusPro cluster, in the top 5 and top 10 largest clusters, and in all clusters (~20 per target). Two of the targets, T0801 and T0851 produced good homodimers for a majority of the CASP11 predicted monomer structures in the top 5, 10, and all ClusPro clusters, while T0770 produced none even though some predictors produced monomer structure models with GDT-TS>60% (Figure 2). Three other targets produced only a few good homodimer models. The RMSD and LDDT results are shown in the Supplementary Material.

We were curious about the very different success rates across the CASP11 targets, and hypothesized that it may have to do with the nature of the oligomers of the available templates for each target. We have developed a database called ProtCID (Protein Common Interface Database),[37] which clusters similar interfaces of homologous proteins (both homodimers and heterodimers) in all of the available crystal forms in the PDB which contain the same Pfam domains.[53] We have found that as the number of crystal forms that contain a particular dimer interface increases, especially if this dimer is in most of the

available crystal forms, it is more likely that the interface is the biologically relevant dimer.[36]

We searched ProtCID for the Pfams of the CASP11 targets, listed in Table I. The numbers of crystal forms of the largest clusters are listed in the table, along with the number of crystal forms available for that Pfam as well as the number of entries in the cluster and in the entire PDB. For instance, targets T0801 and T0843 are in Pfam *DegT_DnrJ_EryC1*, (DegT/DnrJ/EryC1/StrS aminotransferases). This Pfam is found in 23 crystal forms and 45 PDB entries. The largest cluster of similar interfaces (shown in Figure 6A) comprises 22 crystal forms and 44 entries (including T0801 and T0843) and this dimer is annotated as the biological assembly in 40 entries (91%) by the PDB and all 44 entries by PISA. This same dimer is the biological dimer for the two CASP11 targets with this Pfam, T0801 (Figure 6B, PDB: 4PIW) and T0843 (Figure 6C, PDB: 4XAU).

As another example, we compared the interfaces of T0792 (PDB: 5A49), which is not yet in ProtCID, built by the appropriate symmetry operators with the other 5 crystal forms in the PDB that contain the OST-HTH Pfam domain. This interface is found in PDB entry 3RCO (Structural Genomics Consortium, unpublished) and is part of the 24-mer biological assembly produced by PISA for 3RCO. The T0792 crystal interface is also identified as the likely biological assembly by PISA and by the authors who provided the T0792 target to CASP.[50] The ClusPro T0792 dimer with the highest value of Q, the T0792 experimental dimer, and the similar 3RCO dimer are shown in Figures 6D, 6E, and 6F respectively.

Table I lists seven targets (including T0792) for which the homodimer assigned by the authors and/or PISA[45] (T0801, T0776, T0843, T0819, T0851, T0792, T0849) also were contained in ProtCID clusters of homologous proteins. Only 5 of these were ProtCID clusters that contained the majority of all crystal forms (and hence templates) for that target (T0801, T0843, T0819, T0851, T0849). What is notable is that the models of four of these targets (all but T0849) were the most successful in the docking experiment (Figure 5). The reason is likely that the servers producing the models used templates in the PDB that had the same dimer as the experimental CASP structures, and that these modeled structures were in some sense primed to form the same homodimers in ClusPro. Three targets (T0764, T0770, and T0852, shown at the bottom of Table I) contained small clusters in ProtCID, but these clusters were not the same as the homodimers in the crystals of the CASP targets. We examined whether any of the ClusPro clusters contained interfaces similar to the ProtCID clusters for these three targets, and did not find any values of Q above 0.1 (data not shown).

Docking of small ligands to predicted protein structures in CASP11

A small number of the CASP11 structure prediction targets contained biologically relevant compounds in the crystals. A few groups produced models of these ligand/protein complexes, but our goal here is to determine whether the predicted protein structures in CASP11 for each target were suitable for docking of ligands, and to analyze docking success as a function of structure prediction accuracy. We selected one target for this experiment, T0805, a nitroreductase from *M. tuberculosis*, with a bound flavodoxin ligand. Other targets either had covalently bound ligands such as PLP or NAG, very large ligands, or a dimeric ligand (cyclic diguanosine monophosphate), which present complications for docking.

Because in most real-world applications of ligand docking, the binding site of the ligand is obvious from the experimental structure of the templates and the predicted structure of the targets, we built a search space box of 28×18×20 Å covering the binding site as identified in the experimental structure of the CASP11 target. We took the starting structure of the ligand from the experimental structure. We first used the program AutoDock Vina and treated the ligand as flexible as well as 10 of the side chains in the binding site (R17, S18, R20, Y71, A103, L106, W159, T160, T161, L162). We assessed the accuracy of the top 9 scoring docked poses (the default of Autodock Vina) and determined the RMSD to the native ligand after superposition of ligand binding residues. The RMSDs of the best docking (of 9) for each CASP11 model are shown in Figure 7A as a function vs GDT-TS, while all 9 for each CASP11 model are shown in Figure 7B. The best docking of the ligand to the experimental structures results in a poor RMSD of 6.2 Å. Docking to the predicted structures shows a few structures that are a little better than docking to the native, although there is little correlation with the accuracy of the model as measured by GDT-TS.

We also developed a Jaccard index measure by determining the list of protein-residue/ligand-atom contacts in the predicted structure and in the experimental structure and calculating the ratio of the intersection of these lists and the union of these lists (not counting duplicates twice), such that a Jaccard index of 1.0 denotes perfect similarity. The results for this index are shown in Figure 7C and 7D (for top Jaccard index poses and all poses, respectively). Unlike the classical ligand RMSD, one of the nine docking poses to the experimental structure outperformed all of docking simulations to the predicted structures.

Because one important task in docking is to identify specific interactions of the ligand with side chains of the protein, we measured the RMSD of ligand-atom/side-chain distances compared to native. These interactions consist of a salt-bridge of the terminal phosphate group to R17 and an interaction of the aromatic rings of FMN with W159. We refer to the RMSD of these distances with the native structure as scRMSD (specific-contact RMSD). The results for the best scRMSD for each model are shown in Figure 7E and the results for all models are shown in Figure 7F.

The ligand RMSD and the scRMSD are compared in Figures 8A (best scRMSD structure, including the native in red) and 8B (all models and the native structures in red). A total of 101 out of 139 (72.7%) models yielded an scRMSD of less than 1 Å, indicating that while the docked ligands may not superpose extremely well with the experimental structure, some of the important charged and hydrophobic contacts are reproduced. As with the ligand RMSD, the scRMSD of docking to the native structure is not better than docking to many of the models. A comparison of the Jaccard index with the classical ligand RMSD is shown in Figure 8C and 8D. In this case, there is some correlation between the two measurements (R^2 = 0.42 and 0.46 for the best Jaccard-index pose and all poses, respectively). Docking to the experimental structure has higher Jaccard-index scores for comparable RMSDs than most of the models (i.e., most of the red points are above the line in Figure 8D).

To ensure that these relatively poor docking results were not a result of the chosen docking method, we repeated the docking experiment with a second docking program, SwissDock.[40] We employed a flexible ligand approach using a 23×15.5×20.5 Å search box. Prior to

calculating the RMSD, we aligned the backbone heavy atoms of the active site residues (i.e. residues within 5 Å of the ligand in the crystal structure) of each model and the crystal. SwissDock was able to dock the ligand to the crystal structure much more accurately than AutoDock Vina with a ligand RMSD of 0.49 Å. While these docking runs performed much better on the crystal structure, the performance on the models was comparable to the AutoDock Vina results. The best docking RMSD of the ligand to any model was 5.2 Å. The correlation of Jaccard index and ligand RMSD with GDT-TS and Active site backbone RMSD are shown in Figure 9. We see a similar L-shaped pattern that we observed for the homodimer docking (Figure 3), especially with the Active site backbone RMSD (RMSD_i) is used a as the measure of quality of the structure: at high RMSD, all of the docked ligand RMSDs are higher than 8 Å. Then when the Active site backbone RMSD falls below 5 Å, some models achieve better (lower) RMSDs (Figure 9D) and (higher) Jaccard index values (Figure 9B). Still, the docking results to the models are relatively poor, which is not terribly surprising, when taken in the context of the active site residue backbone RMSDs, as the median RMSD_i for all models is 3.8 Å and the best RMSD_i was 2.0 Å. Almost certainly a method that produced an ensemble of conformations of the backbone and the side chains of the active site might perform better.

The experimental ligand structure is shown in each panel of Figure 10 in magenta and the experimental active side chains are in pink. The docking to the experimental structure by AutoDock Vina (Figure 10A) and SwissDock (Figure 10C) are shown in dark blue. The best docking models with AutoDock Vina (Figure 10B) and SwissDock (Figure 10D) show the ligands in dark orange and the model side chains in pale yellow.

Protein structural features used in prediction of missense mutation phenotypes

Another common goal in protein structure prediction is the analysis of missense mutations that arise in the germline or in tumors.[9,17,54] The models may be used to provide plausible mechanisms of action of the known deleterious mutations, or in predictions of the phenotypes by machine-learning algorithms. In the latter case, features that ordinarily may be determined from experimental structures must be obtained from the models, and it is therefore worthwhile to examine the accuracy of predicted features versus the same features calculated from the experimental structure. Here, we focus on a commonly used feature from structure in missense phenotype predictors, the relative accessible surface area of amino acids.[42,43]

There were 7 human proteins among the CASP11 targets and all of these contained one or more domains in the TBM category. We searched for known mutations in these proteins in Uniprot,[55] the Exome Variant Server (EVS),[56] COSMIC,[57] and BioMuta databases.[58] For three of these targets, we were able to find a sufficient number of mutations to analyze. The three targets were T0783 (Uniprot: ISPD_HUMAN; PDB: 4CVH), T0794 (Uniprot: VNN1_HUMAN; PDB: 4CYG), and T0812 (LAMA2_HUMAN), and we were able to identify 27, 20, and 31 germline and somatic missense mutations in these genes respectively. The mutations, sources, surface areas, Polyphen2 predictions,[59] and phenotypes (if available) are listed in Table II.

Since we do not have phenotypes for most of these mutations, it is not possible to compare the predicted CASP models with the experimental structure in terms of phenotype prediction, but we can measure features that might be used in machine-learning predictors. A commonly used feature is the relative solvent accessible surface area (rSASA).[42] We have used the program VMD[60] to calculate the rSASA of the mutations in Table II in both the predicted structures from CASP11 and the experimental structures, and we have calculated the root-mean-square deviation of the predicted values relative to the experimental values (given in Table II). In Figure 11, the RMSDs of rSASA for the relevant mutations for the models of each of the three proteins are graphed vs the accuracy of the predicted structure as measured by GDT-TS (Figure 11 A,B,C). For two of the targets, there was good correlation of the RMSD of the surface area predictions with GDT-TS with R^2 of 0.68 and 0.53 for T0783 and T0794 respectively. However, the correlation is poor for TBM-hard target T0812, where none of the predicted structures have GDT-TS above 45%. Very few of the models have RMSD of rSASA of less than 20% for this target, while many models have RMSD of less than 20% for T0794 and even below 10% for T0783. To confirm that our results were not biased by a restricted sample set, we repeated this procedure for all residues in the proteins, which generated very comparable results (Figure 11 D,E,F).

Another common way to assess the functional effects of mutations is to determine the Gibbs free energy of mutation. We used FoldX[44] to calculate the ΔΔG of mutation for all polymorphisms in the T0812, T0783, and T0794, and we determined the RMSD of ΔΔG for models versus the crystal structures (Figure 11 G,H,I). There were no strong correlations between the similarity of ΔΔG values to those from the crystal structure, except perhaps within the group of models with GDT-TS above 60 for T0794 (Figure 11H).

Missense mutations in the human isoprenoid synthase domain-containing protein (Uniprot ISPD_HUMAN), target T0783 (PDB: 4CVH[61]), are associated with the A,7 form of Muscular Dystrophy-dystroglycanopathy (MDDGA7) or Walker-Warburg syndrome.[62] The missense mutations associated with disease cluster around the active site of the N-terminal domain. A structure alignment of the experimental structure and the best structure prediction of this target is shown in Figure 12A. The positions and even side-chain conformations of the wildtype residues are predicted well by the model, and thus any inference on the functional consequences of mutations of these positions from the model would be highly similar to those from the experimental structure.

The Vanin-1 protein (VNN1) is a membrane-bound and secreted pantotheinase that catalyzes the break down of (R)-pantotheinate into cysteamine and (R)-pantothenate (vitamin B5).[63] Its function is not entirely understood, but it is active under oxidative stress and reduces the levels of reactive oxygen species.[64] As a membrane protein, it may participate in protein-protein interactions that assist precursor T-cells in localizing to the thymus.[63] No missense mutations are known to be associated with disease directly, and there is more evidence for increased activity leading to disease phenotypes through mutations in the promoter region.[65] Nevertheless, some active site mutations might increase activity (or decrease it) as may mutations at some allosteric sites. In Figure 12B, we show the best model produced in CASP11 (by LEER) and the experimental structure superimposed. The magenta sticks are the experimental wild-type residues of the mutations listed in Table II, and the pink sticks

are the side chains of the model. The ligand in spheres is a pantothenate-derived inhibitor. Most of the mutations are predicted directly where they should be (both are numbered when they are separated by more than 3.0 Å), except for some mutations on long external loops like Asp155 and Arg157 and His244.

Heterooligomeric targets in CASP11

Many if not most proteins form physical associations with other proteins (heterooligomeric complexes) in the course of carrying out their functions. We searched protein-protein interaction databases such as STRING[66] and BioGrid,[67] given in the Uniprot pages for each CASP target for possible PPIs that could be modeled for any of the CASP targets. We found 44 CASP11 targets with protein interactions that were assigned high confidence of physical interactions in these databases, and used our program BioAssemblyModeler (BAM)[68] to search the PDB for templates which might contain Pfam domains from the CASP target and its interacting proteins. BAM works by assigning Pfams to a query consisting of up to six different protein sequences, and then searches our PDBfam database[53] for any structures which contain one or more of the Pfams in the query proteins. In the end, we found no such complexes that could be readily modeled without protein docking, primarily because most of the CASP targets are bacterial enzymes and there are few heterooligomeric structures in the PDBs for the associated Pfams.

However, three heterooligomeric structures were provided by experimental groups as targets in CASP11. For technical reasons, these were assigned two target IDs but were evaluated primarily as oligomers and were not included in the regular TBM assessment. These complexes are: 1) T0787/T0788 (PDB: 4TVP[69]), a heterohexameric structure of a trimer of HIV gp120 each bound with gp41; 2) T0797/T0798 (PDB: 4OJK[70]), a heterotetramer of a coiled-coil homodimer from c-AMP-dependent kinase 2 with one copy of human Ras11b bound to each helix; 3) T0840/T0841 (PDB: 4QT8[71]), a heterodimer of the Macrophage-stimulating protein (MSP) receptor and its ligand Hepatocyte growth factor-like protein (also called MSP).

The first oligomeric target, T0787/T0788 (PDB: 4TVP), was a heterohexamer structure of HIV gp120 and gp41, in a closed, prefusion state.[69] The available templates of gp41 in a prefusion state, PDB entries 3J5M[72] and 4NCO,[73] do not contain an alignment of the gp41 sequence to the coordinates since the resolutions were 5.8 Å (cryo-EM) and 4.85 Å respectively. There are also substantially more residues in the coordinates of the CASP target than there are in the templates. The best GDT-TS of gp41 was only 29.0 by the BAKER group with a sequence-independent alignment accuracy (AL0) of only 16%. The gp120 target structure contains an anti-parallel β sheet of the N-terminus and the C-terminus with gp41 wrapped around it, that was not contained in the templates. None of the predictions were able to reproduce this feature (not shown).

Target T0797/T0798 (PDB: 4OJK[70]) is a heterotetramer of a parallel, coiled-coil homodimer from c-AMP-dependent kinase 2 with one copy of human Ras11b bound to each helix. In the PDB, there are two templates of other Ras domains bound to coiled-coil Ras effectors, including a parallel, coiled-coil homodimer of Rab11 family-interacting protein 3 which interacts with two individual copies of Rab11a (PDB: 2HV8[74]) and a parallel, coiled-coil

homodimer of Rab11 family-interacting protein 2 which also interacts with two individual copies of Rab11a (PDB: 2GZD[75]). The two Rab11a structures in the PDB are very similar to each other, as shown in Figure 13A and 13B. Most of the models produced by CASP predictors resemble these templates, shown in Figure 13C.

However, the target Rab11b structure is quite different with the coiled-coil helices interacting with a different surface of the Ras domain. This structure is shown in Figure 13D with the green Ras11b monomer oriented in the same way as the green Ras11a monomers in Figures 13A and 13B. The only successful model of the full tetramer was from the group Seok, who produced an accurate heterotetramer shown in Figure 13E with a GDT-TS of 70, which is a full 20 points better than any other group. The Seok group used docking with the program ZDOCK,[76] and additional loop modeling and refinement to produce these models.

The third heterooligomeric complex in CASP11 was T0840/T0841 (PDB: 4QT8[71]), a heterodimer of the Macrophage-stimulating protein (MSP) receptor and its ligand Hepatocyte growth factor-like protein (also called MSP). The authors suggested that the interest in modeling was to establish the differences between the obvious template, a complex of MET and HGF-β (PDB: 4K3J[77]), and the target structure. The naïve model produced by aligning the target sequences to the template structure and copying the backbone coordinates of aligned residues results in a GDT-TS over the heterodimer of 66.7%, while the best models submitted to CASP of the refined heterodimer had GDT-TS of 70.0%. Models from the Seok and Legato groups do have a shift in the position of two loops of the MSP receptor (after superposition of MSP) from about 5 Å away from the target structure in the naïve model to about 3 Å in the CASP11 models (not shown).

Hypothesis generation

One common use of template-based structure prediction is to generate hypotheses for protein sequence-structure-function relationships. Many of the CASP11 targets are of unknown function and it is difficult to generate hypotheses purely on the basis of structure. Structural information of protein/protein, protein/peptide, protein/ligand, and protein/nucleic-acid interactions from templates can be used to infer possible interaction sites on modeled proteins. An example of this in CASP is target T0856, the SPRY domain of human HERC1 (Uniprot HERC1_HUMAN), an E3 ubiquitin ligase.[78] One of the targets of HERC1 is TSC2, which is inhibited by the interaction of TSC1 with TSC2. SPRY domains have a variety of functions, but in E3 ligases their role appears to be in binding specific substrates.[79]

Our ProtCID database contains not only domain-domain interactions in crystals of homologous domains or domain-pairs, but also a clustering of peptide-domain interactions for each Pfam in the PDB with multiple peptides bound. We looked up the SPRY Pfam domain in ProtCID and found four PDB entries with bound peptides in a single cluster with similar SPRY-domain/peptide binding locations and orientations of the peptides. These include the *Drosophila* Gustavus protein with bound VASA1 peptide (PDB: 2IHS[80]), human SPSB1 (PDB: 3F2O[81]) and SPSB2 (PDB: 3EMW[81]) with bound VASA1 peptide, and human SPSB1 with bound Par4 peptide (PDB: 2JK9[81]). A superposition of these structures with the peptides shown in magenta is shown in Figure 14A. It is unknown what peptides

bind to the SPRY domain of HERC1, but the similarity of binding of peptides to other SPRY domains provides a useful hypothesis of what site on HERC1 may be used for binding peptides in substrate recognition for the E3 ubiquitin ligase activity of HERC1. The best models of HERC1 in CASP11 have accurate alignments of the three binding loops in the interface with the peptides in the ProtCID cluster, as shown in Figure 14B. There are shifts of 3–5 Å between the models and the target structure, but nevertheless the same residues in the SPRY domain are in contact with the peptide in both predicted and experimental structures. Experiments that could be designed from the model to test for interactions of HERC1 with other proteins would be similar to those derived from the experimental structure.

Discussion

Over the history of CASP since 1994, the prediction tasks in the experiment have evolved from secondary structure prediction, fold identification, and accurate sequence alignment in the early CASPs[82–84] to template-based and free modeling,[85] model refinement,[86] and contact-assisted prediction[87] in the most recent CASPs. The emphasis, of course, has been on methods development to produce better and better models in ever more challenging tasks, from comparative modeling all the way through ab initio structure prediction.

Nevertheless, the point of developing methods for protein structure prediction is to apply them to problems in biology and medicine. Predicted structures, just like experimental structures, can be used to design experiments and to understand experimental data and to develop hypotheses on how proteins accomplish their function in complex pathways. Indeed, to measure how widespread the use of structure prediction is in experimental biology, we examined the citation data from Web of Science of five popular comparative modeling platforms, namely Modeller,[88–90] SCWRL,[19,91,92] I-TASSER,[93,94] HHpred,[95] and SwissModel.[96–98] After removing duplicates, we found that these methods were cited a total of 21,396 times. A sampling of 500 of these citing papers suggests that only approximately 5% are methods papers or reviews. Thus, we project that there are over 20,000 manuscripts to date that have used comparative modeling methods to probe real biological questions. Of course, with the many other programs that are available, the true number is much higher.

While the original goal of the CASP organizers was to ask predictors to respond to the biological questions posed by the authors of target structures, only a few such questions were provided by the depositors. In the process of examining several of these, it became evident that we could not assess different predicting groups for their ability to answer such biological questions. We also considered using the CASP11 models in some form of function prediction ourselves, but that is a very challenging task and few of the targets had significant functional information to use in evaluation. We also believe that structural information is more relevant to how proteins carry out biological functions, via binding and changes in structure and dynamics, rather than being useful in predicting function itself. Thus, it made more sense to examine quantifiable predictions of biological function having to do with binding and the effects of mutations that could be obtained from models, and to assess the success of models as a function of the model accuracy as calculated by the standard measures used in CASP. In this way, we could demonstrate for at least a few targets

how structural models might be used in biological research. We chose three such tasks – docking of homodimers, small-ligand docking , and prediction of structural features used in machine learning approaches to missense mutation phenotypes. The advantage we have in CASP is a wide array of accuracies in the ensemble of predicted structures for each target so that we can judge readily how biological function can be predicted depending on the accuracy of the individual model.

In the homodimer docking task, we consistently found an L-shaped pattern of docking accuracy as a function of the accuracy of the monomer models. At low GDT-TS, usually below 60%, none of the monomers formed the correct homodimer in any of the top clusters produced by ClusPro. At some value above 60%, depending on the target, many of the models were able to achieve a good homodimer docking. However, at similar values of GDT-TS, many models were not able to dock the monomers accurately. The same was true of other measures of monomer accuracy such as RMSD and LDDT.

We utilized two docking programs to test the utility of models in small-molecule docking, SwissDock and AutoDock Vina. For AutoDock Vina, the models generated by the CASP11 predictors performed as well as the experimental structure. While few produced impressive ligand RMSDs, most were able to reproduce some of the specific interactions present in the crystal structure. Among all evaluation criteria, there were no strong correlations between model quality and docking results. In this particular case, a limiting factor appears to be the robustness of the docking method, since AutoDock Vina did not dock the ligand to the crystal structure correctly. Conversely, SwissDock was able to dock the ligand very successfully into the crystal structure (0.49 Å RMSD) but the results with the models were similar to the AutoDock Vina results, with the best RMSDs of docking to the model at ~5 Å. The ligand docking results were thus disappointing, demonstrating that ligand docking to models is still a very challenging prospect and not likely to succeed unless the binding site is very accurate (much better than the 2.0 Å backbone RMSD that the best models of target T0805).

In the prediction of surface areas that might be used in missense mutation predictors, we found a linear relationship between RMSD of surface areas from the experimental values versus GDT-TS for two of the targets. The R^2 values were 0.68 and 0.53 for targets T0783 and T0794 respectively. For one other target, the RMSDs did not vary significantly with GDT-TS. We also used the models to calculate ΔΔG of mutation with the program FoldX, but found that there was little correlation with model accuracy, except perhaps within a group of high-accuracy models of one of the targets. This issue needs to be examined further to determine the utility of FoldX for the prediction of missense mutation phenotypes and for the suitability of comparative models for this purpose.

We also examined several targets for protein-protein interactions. Three of the CASP targets were heterooligomers whose interactions were of interest to the crystallographers. For the HIV gp120/gp41 prefusion complex, none of the predictors improved upon the templates significantly. This has to do with the extensive structural changes between the templates and the target structure as well as the fact that gp41 in the templates did not have the sequence aligned to the coordinates. For MSP and its receptor, the models were quite similar to the

templates with a slight improvement in the interaction of loops in the interface. The third target was Ras11b and a homodimer of coiled-coil effector proteins. The templates were quite misleading with a different surface of the Ras domain interacting with the coiled-coil than that in the experimental structure of the target. One group (Seok) was able to use docking to produce a very accurate structure of the heterotetramer.

We showed one example where a knowledge of the molecular interactions in crystal structures of homologues might be used to generate a hypothesis of molecular function of one of the CASP11 targets. The SPRY domain of human HERC1 is likely to be responsible, at least in part, for determining substrates for ubiquitination, including TSC2. Several structures of SPRY domains in the PDB have peptides bound on the same surface of the domain in relatively similar orientations. The models of HERC1, CASP11 target T0856, when superimposed on these domains show a similar interface that is well modeled in the predicted structures compared to the experimental structure of the SPRY domain of HERC1. The best models were not those with the highest GDT-TS, since some of these aligned one of the loops to the template incorrectly. Mutations to disrupt interactions for specific substrates could be easily designed based on the models as readily as they might be on the experimental structure.

In sum, the CASP11 experiment allowed us to perform a number of quantitative assessments of how well comparative models might be utilized in tasks associated with biological function typically performed on experimental structures, including molecular binding and the effects of mutations. Instead of using a large number of different protein targets and one method for producing models, we were able to use many models of individual targets and compare the accuracy of binding and mutation calculations as a function of the accuracy of the model. There is certainly room for additional assessments of this kind, and we hope the data presented in this paper may serve as a pilot experiment of this kind of analysis.

Methods

Docking with ClusPro and assessment of protein-protein docking of CASP11 monomers into homodimers

CASP11 monomers were submitted to the ClusPro server[99] with an application programming interface (API), and the results were retrieved by a PHP script file provided by ClusPro group. Docked protein-protein interfaces were compared to the experimental dimer using the score function Q. The score Q reflects the similarity of contacts between two protein dimers. Because a docked structure may have incorrectly placed side chains, while still having the two proteins in roughly the right orientation and distance from one another, the Q score is appropriate because it depends only the Cβ–Cβ distances of corresponding amino acids in the two interfaces (we use Cα for glycine in what follows). It is a weighted sum of differences in distances of corresponding backbone atoms in the two interfaces. If we define the two distances as e_{ij} and f_{ij} and a weight w_{ij} (a monotonically decreasing function of $d_{ij} = \min\{e_{ij}, f_{ij}\}$), then Q is defined as a sum over all contacts in one or both structures:

$$Q=\frac{\sum_{i,j} w_{\mathrm{ij}} \exp(-0.5|e_{\mathrm{ij}} - f_{\mathrm{ij}}|)}{\sum_{i,j} w_{\mathrm{ij}}}$$

A contact is included in the sum if either e_{ij} or f_{ij} is less than 12 Å. w_{ij} is defined empirically:[36]

$$w_{\mathrm{ij}}=\begin{cases} 1 & 0<d_{\mathrm{ij}} \leq 5 \\ \exp\left(\frac{-(d_{\mathrm{ij}}-5)^2}{9.159}\right) & d>5 \end{cases}.$$

The denominator is the total number of unique contacts in the two structures, where longer contacts between 5 and 12 Å are downweighted exponentially. The distance in the formula for w_{ij} uses the smaller of the Cβ-Cβ distance of the contact in either the predicted or experimental structure. If it is in either structure or both, it will be counted only once. The min operator acts like a logical "or." The numerator is a sum over the same list of contacts that are in the denominator, but if the distance is very different in the predicted structure than in the experimental structure, it is multiplied by a very small number and hence effectively not counted. The only contacts that are counted are those that are similar in the two structures. So Q is approximately the number of correct contacts divided by the total number of unique contacts, or a Jaccard similarity.

The interface residues are defined as these residues with at least one Cβ-Cβ distance ≤ 12 Å between two chains. The LGA program[52] was used to calculate GDT-TS and RMSD of the interface residues by specifying the residue ranges using the parameters –er1 (prediction) and –er2 (target). GDT-TS_i and RMSD_i refer to GDT-TS and RMSD of the interface residues respectively.

Docking of FMN to Target T0805

A structural alignment was performed on all predicted structures and the experimental structure using Theseus.[100,101] For the AutoDock Vina calculations, a 28 × 18 × 20 Å search space was used for the docking. The ligand and ten of the receptor residues (R17, S18, R20, Y71, A103, L106, W159, T160, T161, L162,) were allowed to be flexible in the simulations. We calculated the specific-contact RMSD of the best 9 ligand binding poses for each model, and selected the minimum RMSD. The inter-atomic distances used for the scRMSD calculations were the minimum distance of the FMN terminal phosphate group oxygens to Arg17 side chain guanidinium nitrogens and the minimum distance of the FMN aromatic ring carbons to Trp159 aromatic ring carbons. The Jaccard similarity index was calculated by compiling lists of protein-residue/ligand-atom contacts in the predicted structure and in the experimental structure and calculating the ratio of the intersection of these lists and the union of these lists:

$$J=\frac{(\#\text{of contacts in both structures})}{(\#\text{experimental contacts})+(\#\text{perdicted contacts})-(\#\text{of contacts in both structures})}$$

Contacts were determined by calculating the minimum distance between atomic nuclei in the ligand and surrounding residues. A distance < 4.5 Å was deemed a “contact” and included in the list.

For the SwissDock simulations, we employed a flexible ligand, rigid side-chain approach using a 23×15.5×20.5 Å search box. We aligned the backbone heavy atoms of the active site residues (i.e. residues within 5 Å of the ligand in the crystal structure) of each model and the crystal. We then calculated the RMSD values of ligand heavy atoms and of active site residue backbone atoms.

Surface area calculations for human targets with known missense mutations

Solvent accessible surface areas (SASA) for these residues were computed in VMD[60] using a probe with radius of 1.4 Å. The percent solvent exposed surface area of a given residue side chain is determined relative to the average total surface area of each side chain (solvent exposed area plus buried surface area). RMSD values of relative SASA were calculated with respect to the crystal structure and compared to GDT-TS.

ΔΔG calculations for human targets with known missense mutations

FoldX is an empirical force field that uses the following equation to approximate the free energy (in kcal/mol) of a protein system:

$$\Delta G=\Delta G_{vdw}+\Delta G_{solvH}+\Delta G_{solvP}+\Delta G_{hbond}+\Delta G_{wb}+\Delta G_{el}+\Delta G_{clash}+T\Delta S_{mc}+T\Delta S_{sc}$$

In this equation, the ΔG_{solvH}, ΔG_{solvP}, and G_{wb} terms represent solvent interaction contributions, ΔG_{vdw} accounts for van der Waals contributions, ΔG_{hbond} represents hydrogen bonding contributions, ΔG_{el} accounts for electrostatic contributions, ΔG_{clash} accounts for steric overlaps, and the $T\Delta S_{mc}$ and $T\Delta S_{sc}$ terms account for the entropic costs of fixing the backbone and side chain, respectively. The BuildModel function of FoldX was used to determine the difference in calculated free energies (ΔΔG) of the wild type and mutant proteins. This function optimally rebuilds all side chains around the residue of interest and calculates interaction energies of the residue with its environment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge funding from the NIH: R01 GM84453 to RLD and a Post-doctoral Trainee Fellowship, T32CA009035-38 to PJH. VM is supported by an Elizabeth Knight Patterson Post-doctoral fellowship. MDA and RLD are supported by NIH CA06972 (Fox Chase Cancer Center). We thank Dima Kozakov for providing access to the ClusPro server and Olivier Michielin and VIncent Zoete for providing access to the SwissDock server.

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish R, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958; 181:662–666. [PubMed: 13517261]
2. Blake C, Koenig D, Mair G, North A, Phillips D, Sarma V. Structure of hen egg-white lysozyme A three-dimensional Fourier synthesis at 2 Angstrom resolution. Nature. 1965:757–761. [PubMed: 5891407]
3. Perutz MF, Kendrew JC, Watson HC. Structure and function of haemoglobin. J Mol Biol. 1965; 13:669–678.
4. Browne WJ, North A, Phillips D, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol. 1969; 42:65–86. [PubMed: 5817651]
5. Greer J. Model for haptoglobin heavy chain based upon structural homology. Proceedings of the National Academy of Sciences. 1980; 77:3393–3397.
6. Lustbader J, Arcoleo JP, Birken S, Greer J. Hemoglobin-binding site on haptoglobin probed by selective proteolysis. J Biol Chem. 1983; 258:1227–1234. [PubMed: 6218162]
7. Andersen CBF, Torvund-Jensen M, Nielsen MJ, de Oliveira CLP, Hersleth H-P, Andersen NH, Pedersen JS, Andersen GR, Moestrup SK. Structure of the haptoglobin-haemoglobin complex. Nature. 2012; 489:456–459. [PubMed: 22922649]
8. Zhang YZ, Gould KL, Dunbrack RJ, Cheng H, Roder H, Golemis EA. The evolutionarily conserved Dim1 protein defines a novel branch of the thioredoxin fold superfamily. Physiol Genomics. 1999; 1:109–118. [PubMed: 11015569]
9. Shan X, Dunbrack RL Jr, Christopher SA, Kruger WD. Mutations in the regulatory domain of cystathionine beta-synthase can functionally suppress patient-derived mutations in cis. Hum Mol Genet. 2001; 10:635–643. [PubMed: 11230183]
10. Cheng JD, Dunbrack RL Jr, Valianou M, Rogatko A, Alpaugh RK, Weiner LM. Promotion of tumor growth by murine fibroblast activation protein, a serine protease, in an animal model. Cancer Res. 2002; 62:4767–4772. [PubMed: 12183436]
11. Kundrat L, Martins J, Stith L, Dunbrack RL Jr, Jaffe EK. A structural basis for half-of-the-sites metal binding revealed in Drosophila melanogaster porphobilinogen synthase. J Biol Chem. 2003; 278:31325–31330. [PubMed: 12794073]
12. Zhang R, Poustovoitov MV, Ye X, Santos HA, Chen W, Daganzo SM, Erzberger JP, Serebriiskii IG, Canutescu AA, Dunbrack RL, Pehrson JR, Berger JM, Kaufman PD, Adams PD. Formation of MacroH2A–containing senescence-associated heterochromatin foci and senescence driven by ASF1a and HIRA. Dev Cell. 2005; 8:19–30. [PubMed: 15621527]
13. Tang Y, Poustovoitov MV, Zhao K, Garfinkel M, Canutescu A, Dunbrack R, Adams PD, Marmorstein R. Structure of a human ASF1a–HIRA complex and insights into specificity of histone chaperone complex assembly. Nat Struct Mol Biol. 2006; 13:921–929. [PubMed: 16980972]
14. Li C, Andrake M, Dunbrack R, Enders GH. A bifunctional regulatory element in human somatic Wee1 mediates cyclin A/Cdk2 binding and Crm1-dependent nuclear export. Mol Cell Biol. 2010; 30:116–130. [PubMed: 19858290]
15. Plotnikova OV, Pugacheva EN, Dunbrack RL, Golemis EA. Rapid calcium-dependent activation of Aurora-A kinase. Nature communications. 2010; 1
16. Tong X, Zitserman D, Serebriiskii I, Andrake M, Dunbrack R, Roegiers F. Numb independently antagonizes Sanpodo membrane targeting and Notch signaling in Drosophila sensory organ precursor cells. Mol Biol Cell. 2010; 21:802–810. [PubMed: 20053677]
17. Roberts JL, Buckley RH, Luo B, Pei J, Lapidus A, Peri S, Wei Q, Shin J, Parrott RE, Dunbrack RL, Testa JR, Zhong X-P, Wiest DL. CD45-deficient severe combined immunodeficiency caused by uniparental disomy. Proceedings of the National Academy of Sciences. 2012; 109:10456–10461.
18. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005; 353:459–473. [PubMed: 16169011]

19. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003; 12:2001–2014. [PubMed: 12930999]
20. Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014; 426:2692–2701. [PubMed: 24810707]
21. Tovchigrechko A, Wells CA, Vakser IA. Docking of protein models. Protein Sci. 2002; 11:1888–1896. [PubMed: 12142443]
22. Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models docking benchmark 2. Proteins: Structure, Function, and Bioinformatics. 2015; 83:891–897.
23. Kaufmann KW, Meiler J. Using RosettaLigand for small molecule docking into comparative models. PLOS ONE. 2012; 7:e50769. [PubMed: 23239984]
24. Bordogna A, Pandini A, Bonati L. Predicting the accuracy of protein-ligand docking on homology models. J Comput Chem. 2011; 32:81–98. [PubMed: 20607693]
25. Fan H, Irwin JJ, Webb BM, Klebe G, Shoichet BK, Sali A. Molecular docking screens using comparative models of proteins. Journal of chemical information and modeling. 2009; 49:2512–2527. [PubMed: 19845314]
26. Ferrara P, Jacoby E. Evaluation of the utility of homology models in high throughput docking. J Mol Model. 2007; 13:897–905. [PubMed: 17487515]
27. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. Drug Discov Today. 2009; 14:676–683. [PubMed: 19422931]
28. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. Drug Discov Today. 2004; 9:659–669. [PubMed: 15279849]
29. Soro S, Tramontano A. The prediction of protein function at CASP6. Proteins: Structure, Function, and Bioinformatics. 2005; 61:201–213.
30. Lopez G, Rojas A, Tress M, Valencia A. Assessment of predictions submitted for the CASP7 function prediction category. Proteins: Structure, Function, and Bioinformatics. 2007; 69:165–174.
31. Lopez G, Ezkurdia I, Tress ML. Assessment of ligand binding residue predictions in CASP8. Proteins: Structure, Function, and Bioinformatics. 2009; 77:138–146.
32. Schmidt T, Haas J, Cassarino TG, Schwede T. Assessment of ligand-binding residue predictions in CASP9. Proteins: Structure, Function, and Bioinformatics. 2011; 79:126–136.
33. Gallo Cassarino T, Bordoli L, Schwede T. Assessment of ligand binding site predictions in CASP10. Proteins: Structure, Function, and Bioinformatics. 2014; 82:154–163.
34. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. Proteins: Structure, Function, and Bioinformatics. 2006; 65:392–406.
35. Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, Vajda S. How good is automated protein docking? Proteins: Structure, Function, and Bioinformatics. 2013; 81:2159–2166.
36. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL Jr. Statistical analysis of interface similarity in crystals of homologous proteins. J Mol Biol. 2008; 381:487–507. [PubMed: 18599072]
37. Xu Q, Dunbrack RL Jr. The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res. 2011; 39:D761–D770. [PubMed: 21036862]
38. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bull Soc Vaud Sci Nat. 1901; 37:547–549.
39. Jaccard P. The distribution of the flora in the alpine zone 1. New Phytol. 1912; 11:37–50.
40. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. 2011; 39:W270–W277. [PubMed: 21624888]
41. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010; 31:455–461. [PubMed: 19499576]
42. Wei Q, Xu Q, Dunbrack RL. Prediction of phenotypes of missense mutations in human proteins from biological assemblies. Proteins: Structure, Function, and Bioinformatics. 2013; 81:199–213.

43. Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLOS ONE. 2013; 8:e67863. [PubMed: 23874456]
44. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005; 33:W382–W388. [PubMed: 15980494]
45. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of molecular biology. 2007; 372:774–797. [PubMed: 17681537]
46. Jin J, Xie X, Chen C, Park JG, Stark C, James DA, Olhovsky M, Linding R, Mao Y, Pawson T. Eukaryotic protein domains as functional units of cellular evolution. Science signaling. 2009; 2:ra76–ra76. [PubMed: 19934434]
47. Vacic V, Iakoucheva LM, Lonardi S, Radivojac P. Graphlet kernels for prediction of functional residues in protein structures. J Comput Biol. 2010; 17:55–72. [PubMed: 20078397]
48. Krause A, Stoye J, Vingron M. Large scale hierarchical clustering of protein sequences. BMC Bioinformatics. 2005; 6:15. [PubMed: 15663796]
49. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol. 2002; 315:1257–1275. [PubMed: 11827492]
50. Jeske M, Bordi M, Glatt S, Müller S, Rybin V, Müller CW, Ephrussi A. The crystal structure of the Drosophila germline inducer Oskar identifies two domains with distinct Vasa helicase-and RNA-binding activities. Cell reports. 2015; 12:587–598. [PubMed: 26190108]
51. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013; 29:2722–2728. [PubMed: 23986568]
52. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31:3370–3374. [PubMed: 12824330]
53. Xu Q, Dunbrack RL Jr. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. Bioinformatics. 2012; 28:2763–2772. [PubMed: 22942020]
54. Sauder JM, Arthur JW, Dunbrack RL Jr. Modeling of substrate specificity of the Alzheimer's disease amyloid precursor protein beta-secretase. J Mol Biol. 2000; 300:241–248. [PubMed: 10873463]
55. Magrane M. UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011; 2011:bar009. [PubMed: 21447597]
56. Project NGES. Exome variant server. Seattle, WA: 2015. URL: http://evsgswashingtonedu/EVS/) 05/2015
57. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43:D805–D811. [PubMed: 25355519]
58. Wu T-J, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, Mazumder R. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database. 2014; 2014:bau022. [PubMed: 24667251]
59. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
60. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graphics. 1996; 14:33–38.
61. Riemersma M, Froese DS, van Tol W, Engelke UF, Kopec J, van Scherpenzeel M, Ashikov A, Krojer T, von Delft F, Tessari M, Buczkowska A, Swiezewska E, Jae LT, Brummelkamp TR, Manya H, Endo T, Bokhoven Hv, Yue WW, Lefeber DJ. Human ISPD Is a Cytidyltransferase Required for Dystroglycan O-Mannosylation. Chem Biol. 2015; 22:1643–1652. [PubMed: 26687144]
62. Willer T, Lee H, Lommel M, Yoshida-Moriguchi T, de Bernabe DBV, Venzke D, Cirak S, Schachter H, Vajsar J, Voit T, Muntoni F, Loder AS, Dobyns WB, Winder TL, Strahl S, Mathews KD, Nelson SF, Moore SA, Campbell KP. ISPD loss-of-function mutations disrupt dystroglycan

O-mannosylation and cause Walker-Warburg syndrome. Nat Genet. 2012; 44:575–580. [PubMed: 22522420]

63. Martin F, Malergue F, Pitari G, Philippe JM, Philips S, Chabret C, Granjeaud S, Mattei MG, Mungall AJ, Naquet P, Galland F. Vanin genes are clustered (human 6q22-24 and mouse 10A2B1) and encode isoforms of pantetheinase ectoenzymes. Immunogenetics. 2001; 53:296–306. [PubMed: 11491533]
64. Ferreira DW, Naquet P, Manautou JE. Influence of Vanin-1 and Catalytic Products in Liver During Normal and Oxidative Stress Conditions. Curr Med Chem. 2015; 22:2407–2416. [PubMed: 26549544]
65. Kaskow BJ, Diepeveen LA, Proffitt JM, Rea AJ, Ulgiati D, Blangero J, Moses EK, Abraham LJ. Molecular prioritization strategies to identify functional genetic variants in the cardiovascular disease-associated expression QTL Vanin-1. Eur J Hum Genet. 2014; 22:688–695. [PubMed: 24045843]
66. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011; 39:D561–D568. [PubMed: 21045058]
67. Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2014; 43(Database issue):D470–D487. [PubMed: 25428363]
68. Shapovalov MV, Wang Q, Xu Q, Andrake M, Dunbrack RL Jr. Bioassemblymodeler (BAM): User-friendly homology modeling of protein homo-and heterooligomers. PLOS ONE. 2014; 9:e98309. [PubMed: 24922057]
69. Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, Huang J, Acharya P, Chuang G-Y, Ofek G, Stewart-Jones GBE, Stuckey J, Bailer RT, Joyce MG, Louder MK, Tumba N, Yang Y, Zhang B, Cohen MS, Haynes BF, Mascola JR, Morris L, Munro JB, Blanchard SC, Mothes W, Connors M, Kwong PD. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. Nature. 2014; 514:455–461. [PubMed: 25296255]
70. Reger AS, Yang MP, Koide-Yoshida S, Guo E, Mehta S, Yuasa K, Liu A, Casteel DE, Kim C. Crystal Structure of the cGMP-dependent Protein Kinase II Leucine Zipper and Rab11b Protein Complex Reveals Molecular Details of G-kinase-specific Interactions. J Biol Chem. 2014; 289:25393–25403. [PubMed: 25070890]
71. Chao KL, Gorlatova NV, Eisenstein E, Herzberg O. Structural Basis for the Binding Specificity of Human Recepteur d'Origine Nantais (RON) Receptor Tyrosine Kinase to Macrophage-stimulating Protein. J Biol Chem. 2014; 289:29948–29960. [PubMed: 25193665]
72. Lyumkis D, Julien J-P, de Val N, Cupo A, Potter CS, Klasse P-J, Burton DR, Sanders RW, Moore JP, Carragher B, Wilson IA, Ward AB. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. Science. 2013; 342:1484–1490. [PubMed: 24179160]
73. Julien J-P, Cupo A, Sok D, Stanfield RL, Lyumkis D, Deller MC, Klasse P-J, Burton DR, Sanders RW, Moore JP, Ward AB, Wilson IA. Crystal structure of a soluble cleaved HIV-1 envelope trimer. Science. 2013; 342:1477–1483. [PubMed: 24179159]
74. Eathiraj S, Mishra A, Prekeris R, Lambright DG. Structural basis for Rab11-mediated recruitment of FIP3 to recycling endosomes. J Mol Biol. 2006; 364:121–135. [PubMed: 17007872]
75. Jagoe WN, Lindsay AJ, Read RJ, McCoy AJ, McCaffrey MW, Khan AR. Crystal structure of rab11 in complex with rab11 family interacting protein 2. Structure. 2006; 14:1273–1283. [PubMed: 16905101]
76. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins. 2003; 52:80–87. [PubMed: 12784371]
77. Stamos J, Lazarus RA, Yao X, Kirchhofer D, Wiesmann C. Crystal structure of the HGF β-chain in complex with the Sema domain of the Met receptor. EMBO J. 2004; 23:2325–2335. [PubMed: 15167892]

78. Chong-Kopera H, Inoki K, Li Y, Zhu T, Garcia-Gonzalo FR, Rosa JL, Guan K-L. TSC1 stabilizes TSC2 by inhibiting the interaction between TSC2 and the HERC1 ubiquitin ligase. J Biol Chem. 2006; 281:8313–8316. [PubMed: 16464865]

79. Perfetto L, Gherardini PF, Davey NE, Diella F, Helmer-Citterich M, Cesareni G. Exploring the diversity of SPRY/B30 2-mediated interactions. Trends Biochem Sci. 2013; 38:38–46. [PubMed: 23164942]

80. Woo J-S, Suh H-Y, Park S-Y, Oh B-H. Structural basis for protein recognition by B30 2/SPRY domains. Mol Cell. 2006; 24:967–976. [PubMed: 17189197]

81. Filippakopoulos P, Low A, Sharpe TD, Uppenberg J, Yao S, Kuang Z, Savitsky P, Lewis RS, Nicholson SE, Norton RS, Bullock AN. Structural basis for Par-4 recognition by the SPRY domain-and SOCS box-containing proteins SPSB1, SPSB2, and SPSB4. J Mol Biol. 2010; 401:389–402. [PubMed: 20561531]

82. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins: Structure, Function and Genetics. 1997; (Suppl): 2–6.

83. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins: Structure, Function and Genetics. 1999; (Suppl):2–6.

84. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins. 2001; (Suppl 5):2–7. [PubMed: 11835476]

85. Taylor TJ, Tai CH, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee B. Definition and classification of evaluation units for CASP10. Proteins. 2014; (82 Suppl 2):14–25. [PubMed: 24123179]

86. Nugent T, Cozzetto D, Jones DT. Evaluation of predictions in the CASP10 model refinement category. Proteins: Structure, Function, and Bioinformatics. 2014; 82:98–111.

87. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact prediction in CASP10. Proteins. 2014; (82 Suppl 2):138–153. [PubMed: 23760879]

88. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234:779–815. [PubMed: 8254673]

89. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins: Structure, Function and Genetics. 1997; (Suppl):50–58.

90. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 2006; 34:D291–D295. [PubMed: 16381869]

91. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol. 1997; 267:1268–1282. [PubMed: 9150411]

92. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009; 77:778–795. [PubMed: 19603484]

93. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins: Structure, Function and Genetics. 2005; 61(Suppl 7):91–98.

94. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9:40. [PubMed: 18215316]

95. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33:W244–W248. [PubMed: 15980461]

96. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997; 18:2714–2723. [PubMed: 9504803]

97. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003; 31:3381–3385. [PubMed: 12824332]

98. Bordoli L, Schwede T. Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal. Methods Mol Biol. 2012; 857:107–136. [PubMed: 22323219]

99. Kozakov D, Hall DR, Beglov D, Brenke R, Comeau SR, Shen Y, Li K, Zheng J, Vakili P, Paschalidis I, Vajda S. Achieving reliability and high accuracy in automated protein docking:

ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19. Proteins. 2010; 78:3124–3130. [PubMed: 20818657]

100. Theobald DL, Steindel PA. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics. 2012; 28:1972–1979. [PubMed: 22543369]

101. Theobald DL, Wuttke DS. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics. 2006; 22:2171–2172. [PubMed: 16777907]

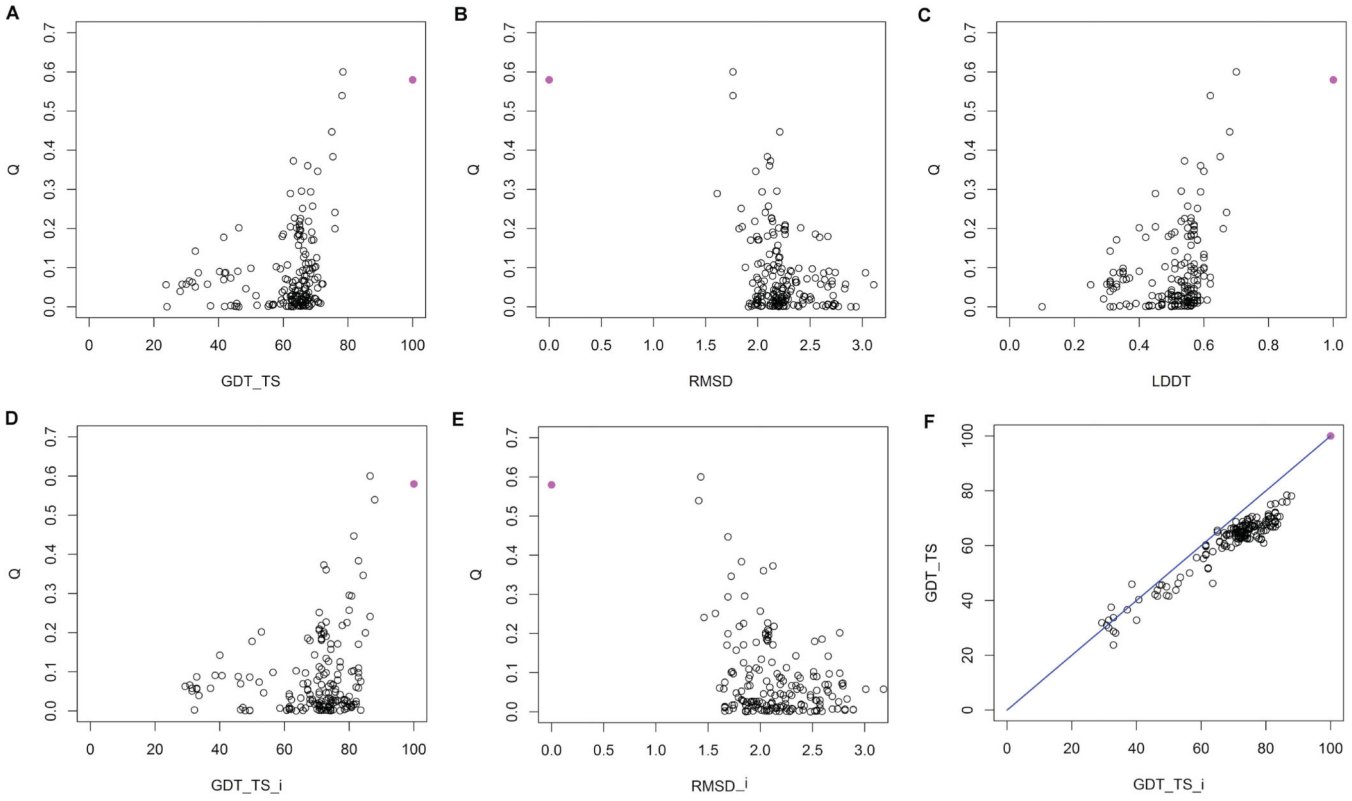

Figure 1.

Q scores of docked homodimers from monomeric models of target T0792. Each predicted monomer structure was submitted to the ClusPro server, and the best *Q* score of the top 10 clusters produced by ClusPro was plotted versus a measure of the model quality: A) GDT-TS; B) RMSD; C) LDDT; D) GDT-TS_i (GDT-TS of the interface residues); E) RMSD_i (RMSD of the interface residues); F) scatter plot of GDT-TS vs. GDT-TS_i. The "_i" refers to the interface residues of the crystal dimer that have at least one inter-chain Cβ-Cβ distance ≤ 12 Å. GDT-TS_i and RMSD_i are calculated from interface residues only by LGA program.

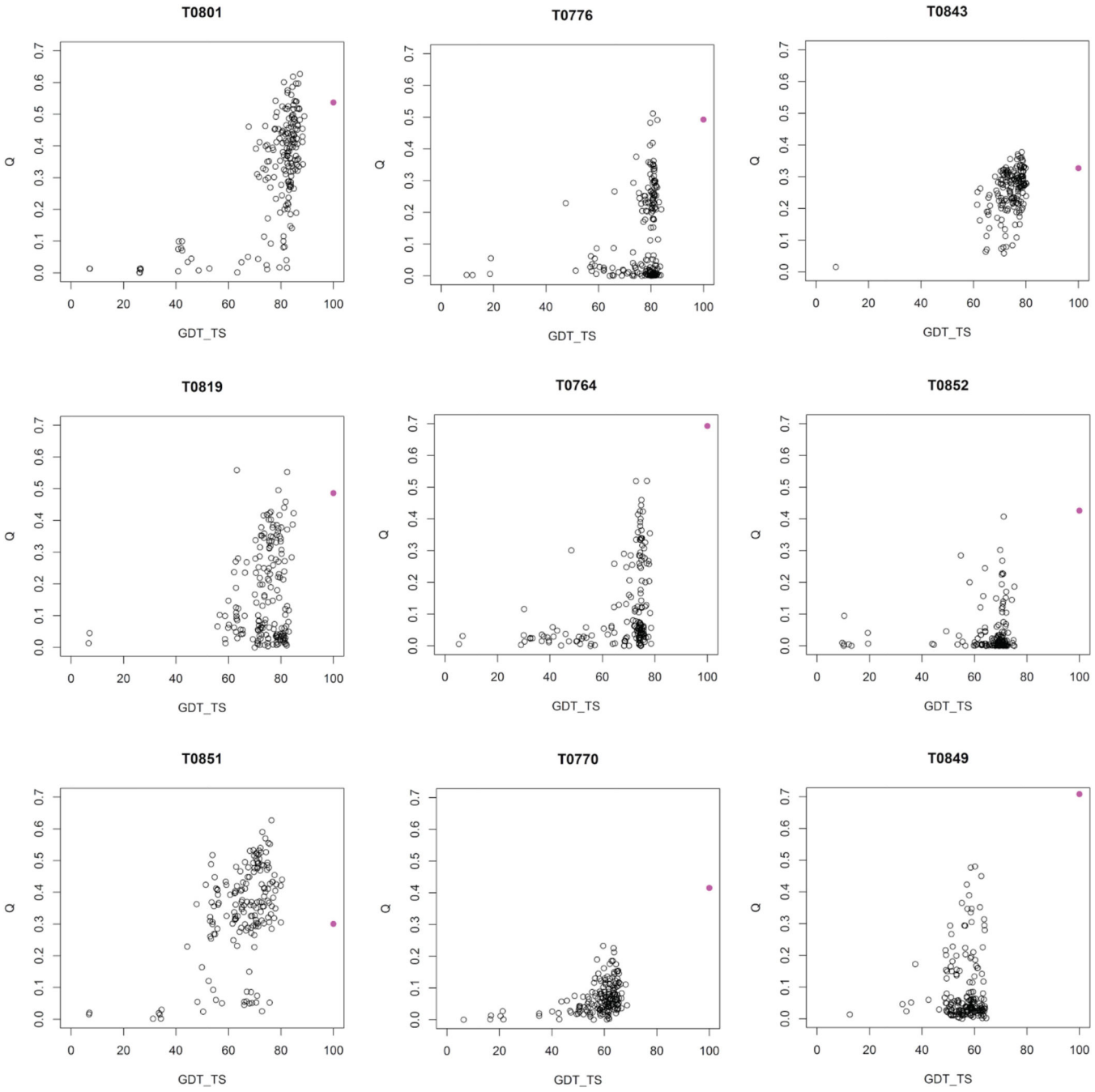

Figure 2.

Q scores of docked homodimers for all targets (except T0792 shown in Figure 1A) vs GDT-TS. The best *Q* score of the top 10 largest ClusPro clusters is shown. The targets are ordered by decreasing average GDT-TS scores (Table I). The same order is used in Figure 3, 4 and 5.

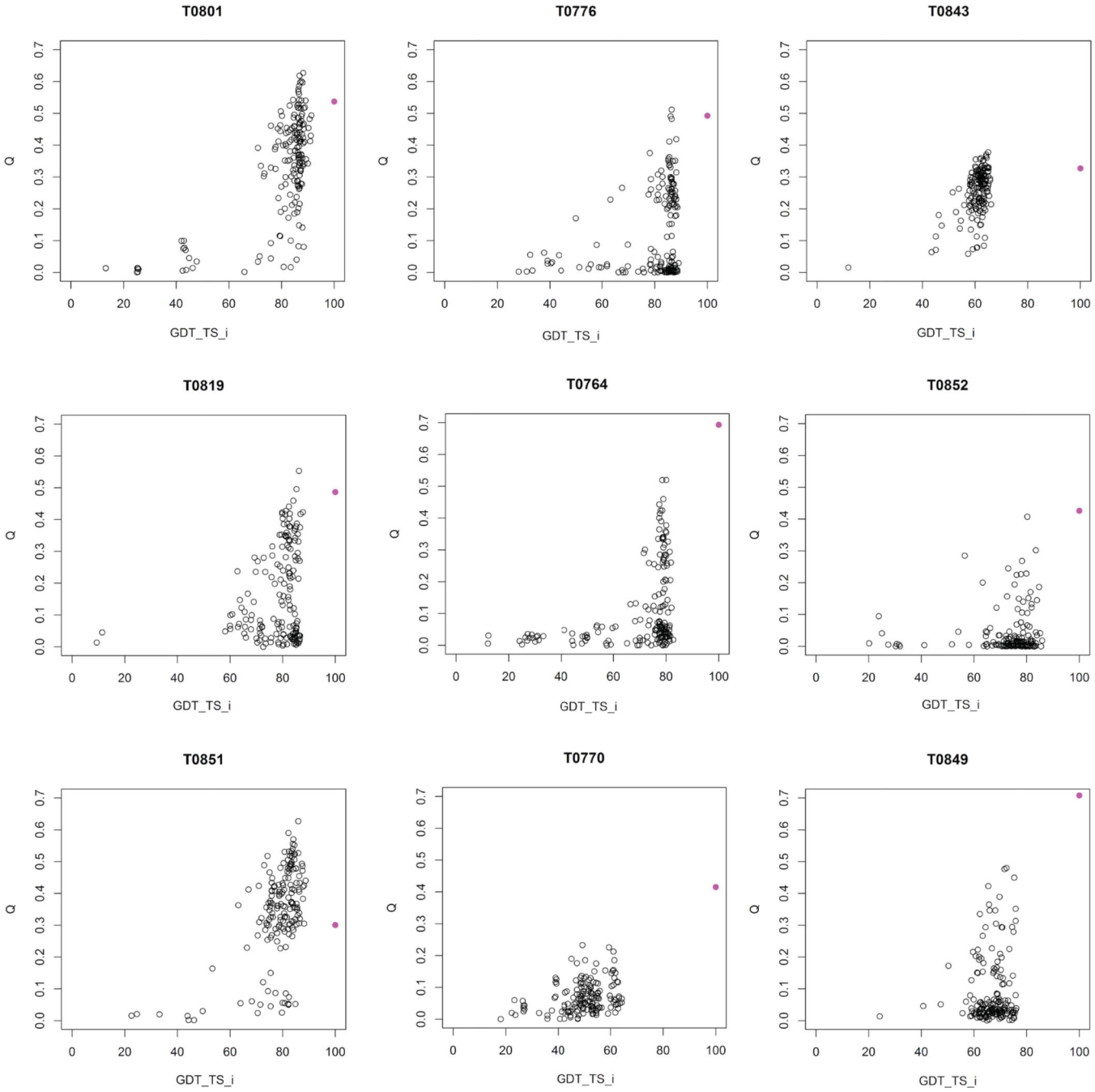

Figure 3.

Q scores of docked homodimers for all targets (except T0792 shown in Figure 1D) vs GDT-TS_i. The best *Q* score of the top 10 largest ClusPro clusters is shown. The targets are ordered by decreasing average GDT-TS scores (Table I)

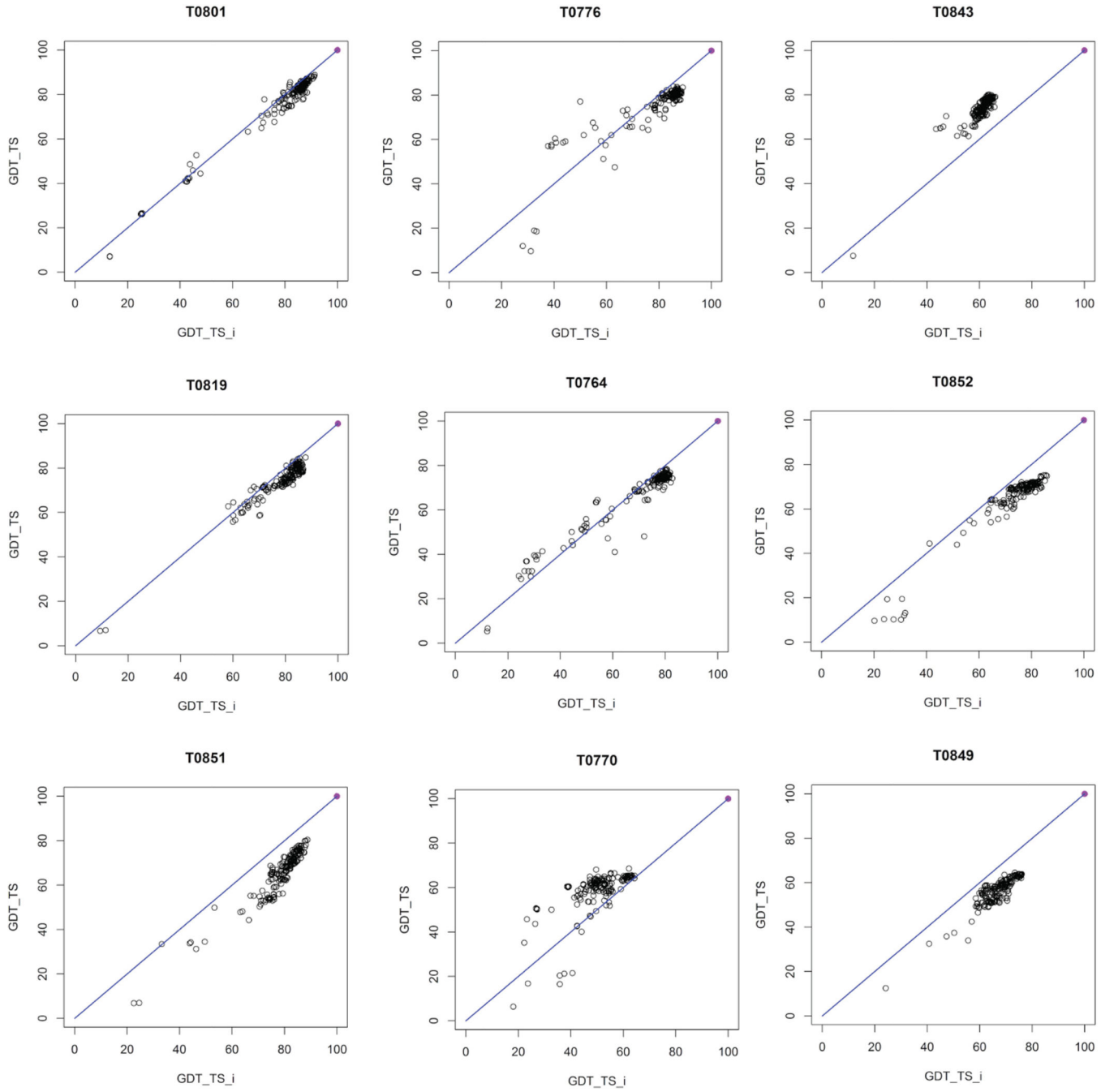

Figure 4.

The scatter plots of GDT-TS and GDT-TS_i for all targets (except T0792 shown in Figure 1F). The targets are ordered by decreasing average GDT-TS scores (Table I)

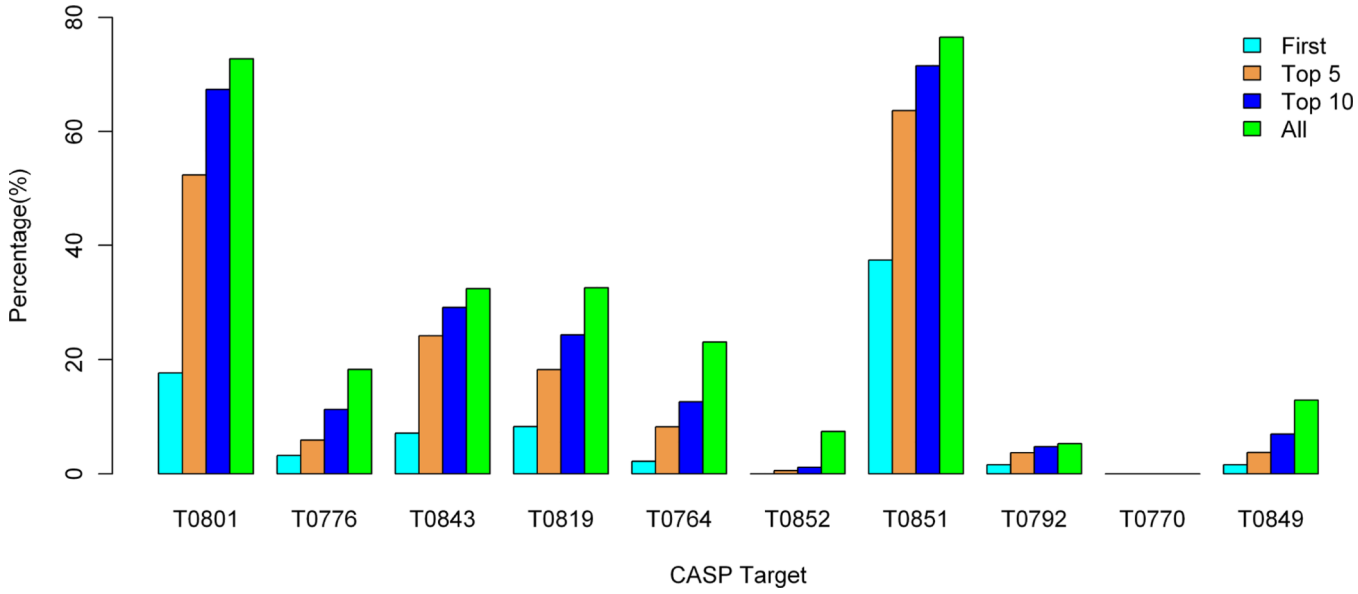

Figure 5.

Percentage of monomer models that were able to achieve a Q score of at least 0.3 for the first cluster, the top 5 clusters, the top 10 clusters, and all clusters (on average ~20 per model). The targets are ordered by decreasing average GDT-TS scores (Table I)

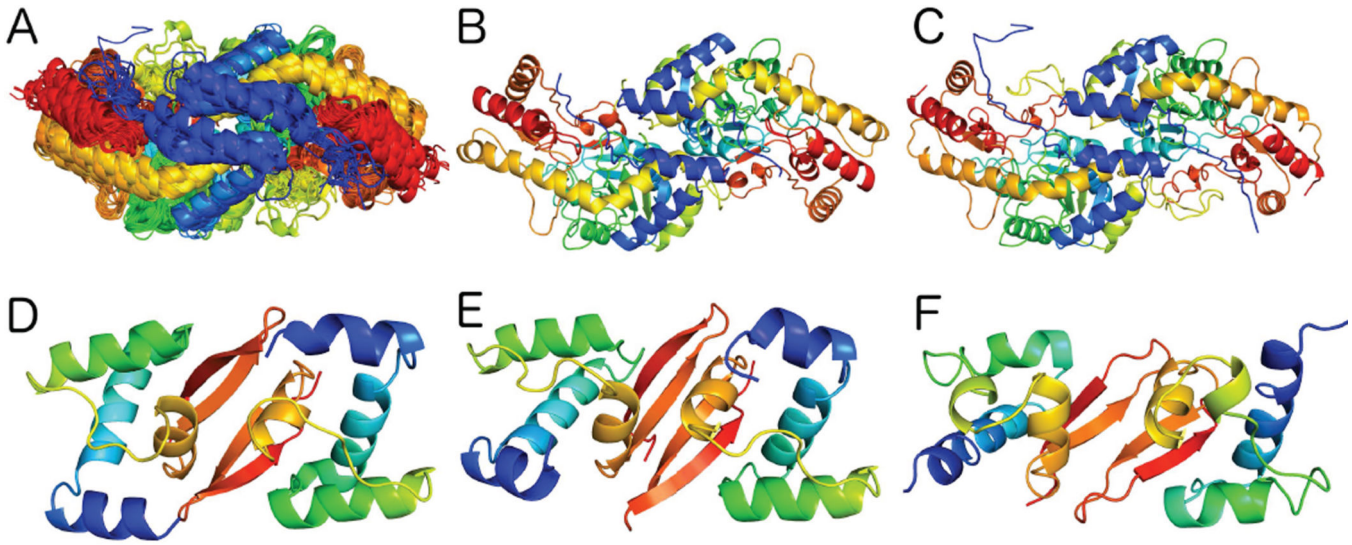

Figure 6.

A. Cluster of similar interfaces from crystals of proteins containing Pfam domain *DegT_DnrJ_EryC1*, (DegT/DnrJ/EryC1/StrS aminotransferases, provided by ProtCID. The interface is observed in 23 crystal forms of this Pfam and 44 PDB entries. B. The biological dimer of CASP11 target T0801 (PDB: 4PIW). C. The biological dimer of CASP11 target T0843 (PDB: 4XAU). D. Predicted structure of T0792 (group 216) docked by ClusPro. E. Experimental structure of T0792. F. Similar dimer in crystal of PDB entry 3RCO (human TDRD7).

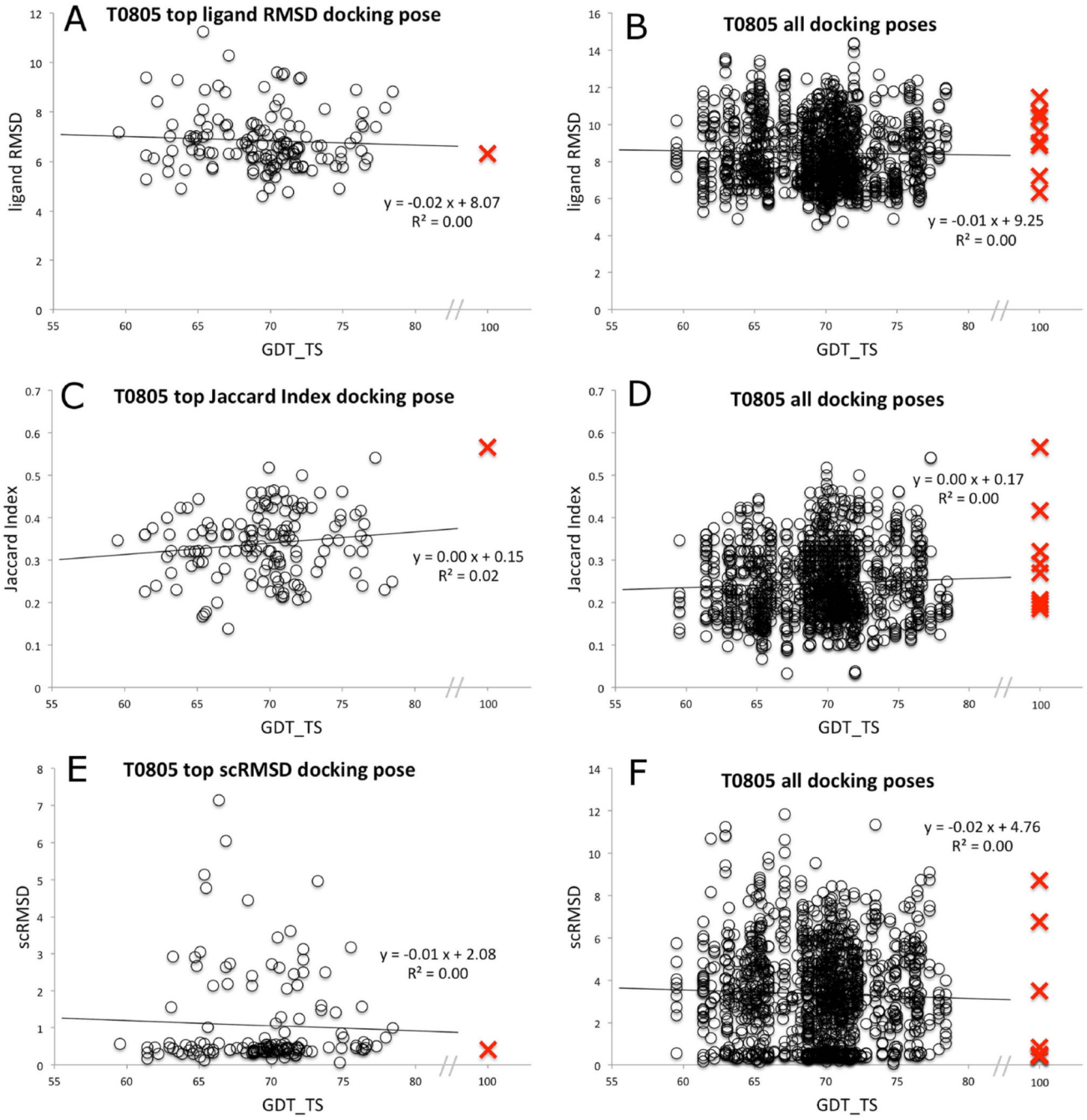

Figure 7.

Docking of flavodoxin to models of target T0805 with AutoDock Vina and comparison with experimental structure. A. RMSD vs. GDT-TS of model for lowest RMSD of 9 docked poses by AutoDock/VINA. B. RMSD vs GDT-TS for all docked poses to each model; C. Jaccard index vs GDT-TS for highest Jaccard index of 9 docked poses. D. Jaccard index vs GDT-TS for all docked poses for each model. E. scRMSD (specific-contact RMSD) vs GDT-TS for top scRMSD docked poses; F. scRMSD vs GDT-TS for all docked poses. For all plots, the results for the experimental structure are marked with red X's.

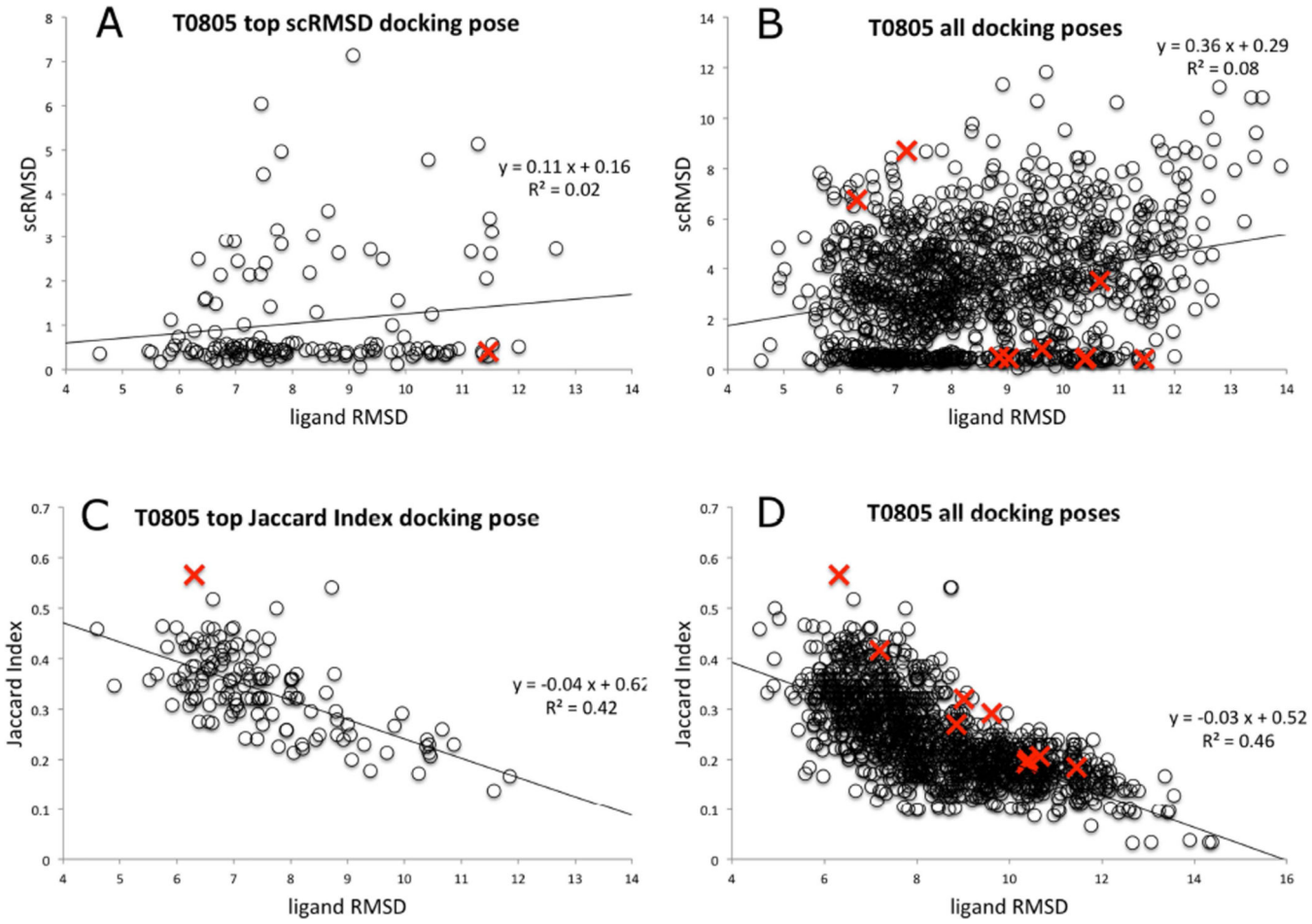

Figure 8.

A. scRMSD vs RMSD for top scRMSD docked poses. B. scRMSD vs RMSD for all docked poses. C. Jaccard index vs RMSD for top scRMSD docked poses; D. Jaccard index vs RMSD for all docked poses. Red X's indicate docking to experimental structure.

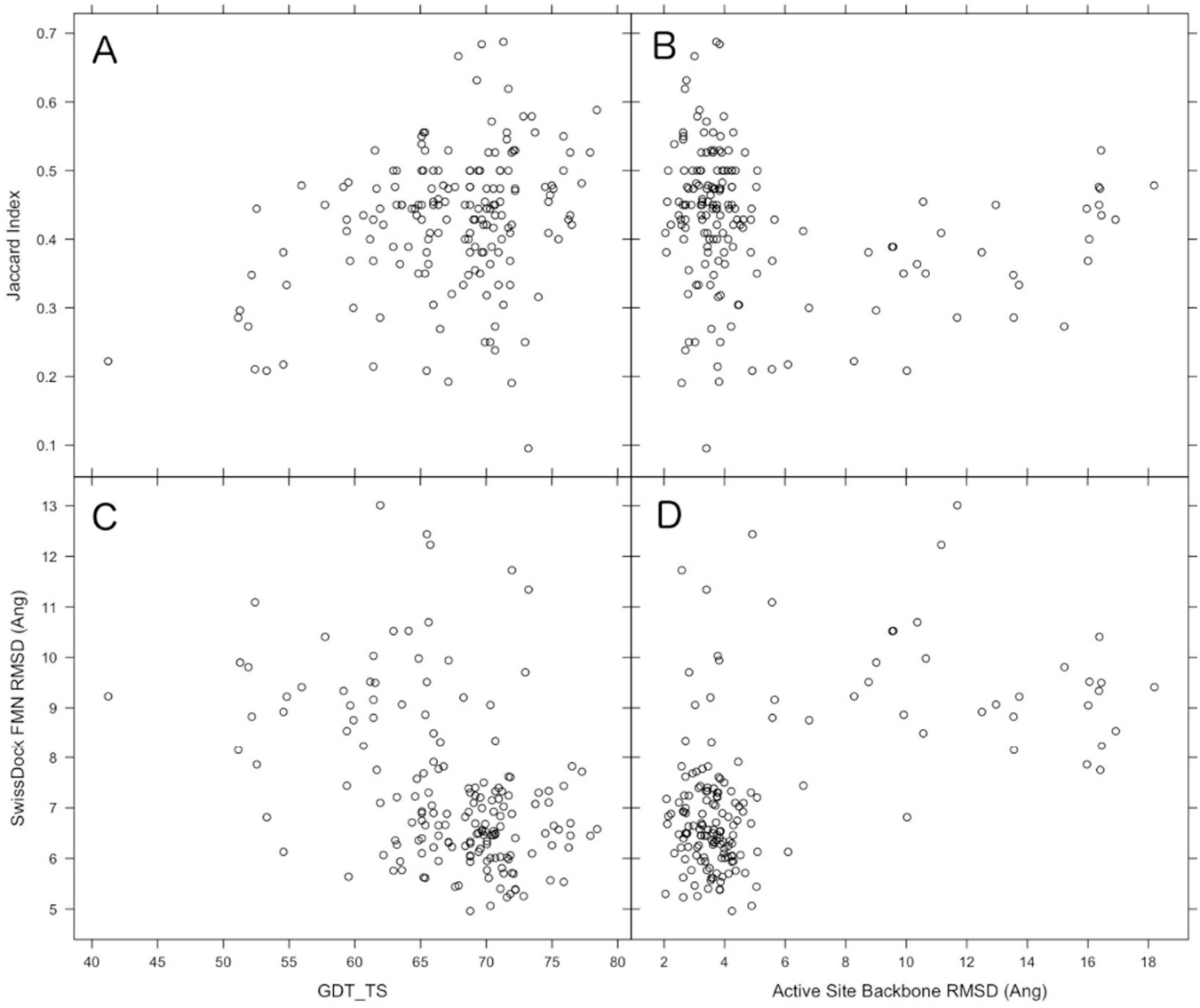

Figure 9.

Docking of flavodoxin to models of target T0805 with SwissDock and comparison with experimental structure. A. Jaccard index vs GDT-TS; B. Jaccard index vs RMSD_i (interface backbone atom RMSD); C. RMSD of ligand vs GDT-TS; D. RMSD of ligand vs RMSD_i. Docking to the experimental structure produced a Jaccard index of 0.93 and a ligand RMSD of 0.49 Å (not shown on plots).

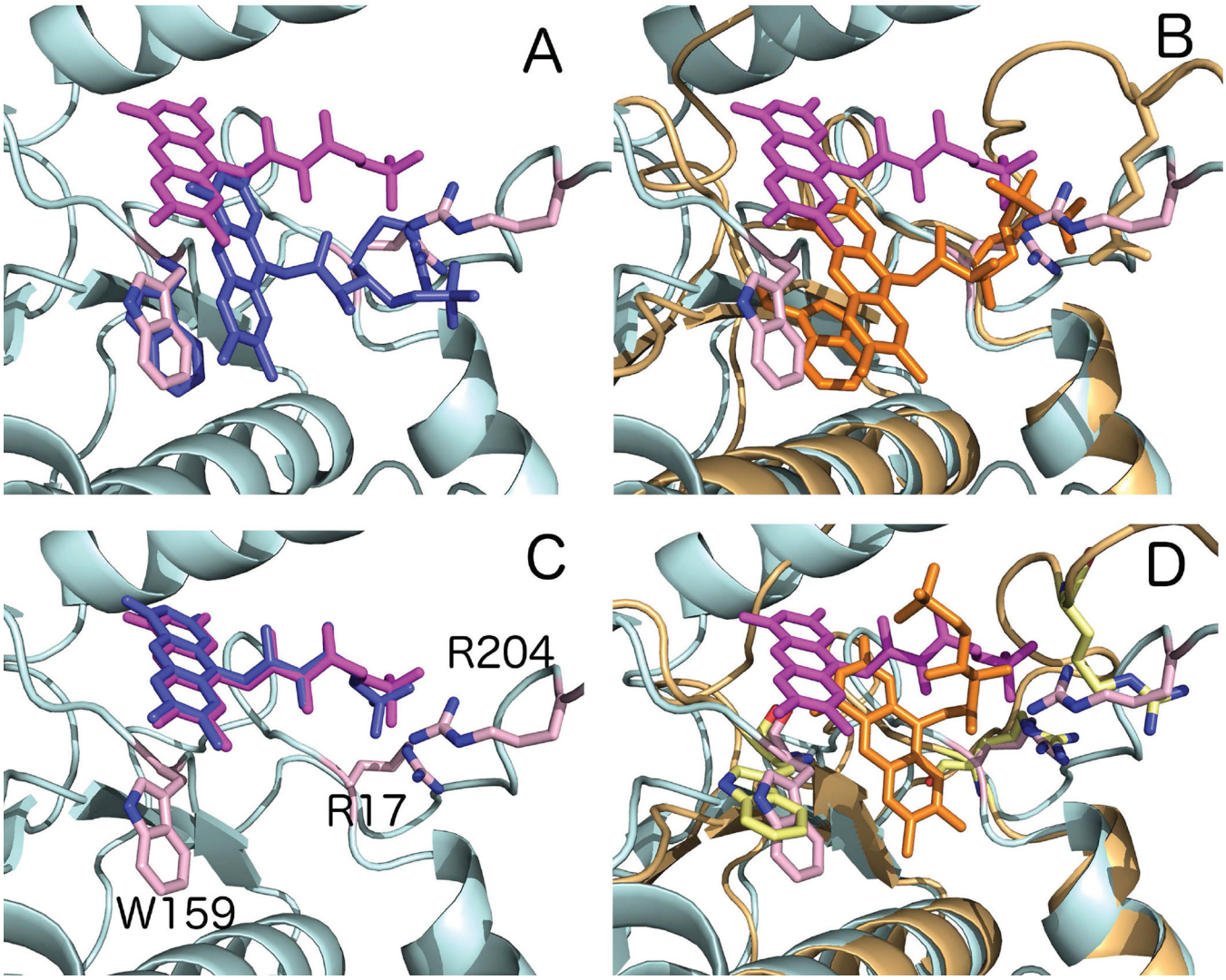

Figure 10.

T0805 docking. A. AutoDock Vina docking of ligand to experimental structure (blue) along with experimental position of ligand (magenta). B. AutoDock Vina docking of ligand to predicted structure by group IntFold3 (model 1) (orange) along with experimental position of ligand (magenta). C. SwissDock docking of ligand to experimental structure (blue) along with experimental position of ligand (magenta); D. SwissDock docking of ligand to predicted structure by group BioSerf (model 2) (orange) along with experimental position of ligand (magenta). Side chains of experimental structure are depicted in pink and labeled in panel C. Side chains of modeled structure are in yellow (panel D).

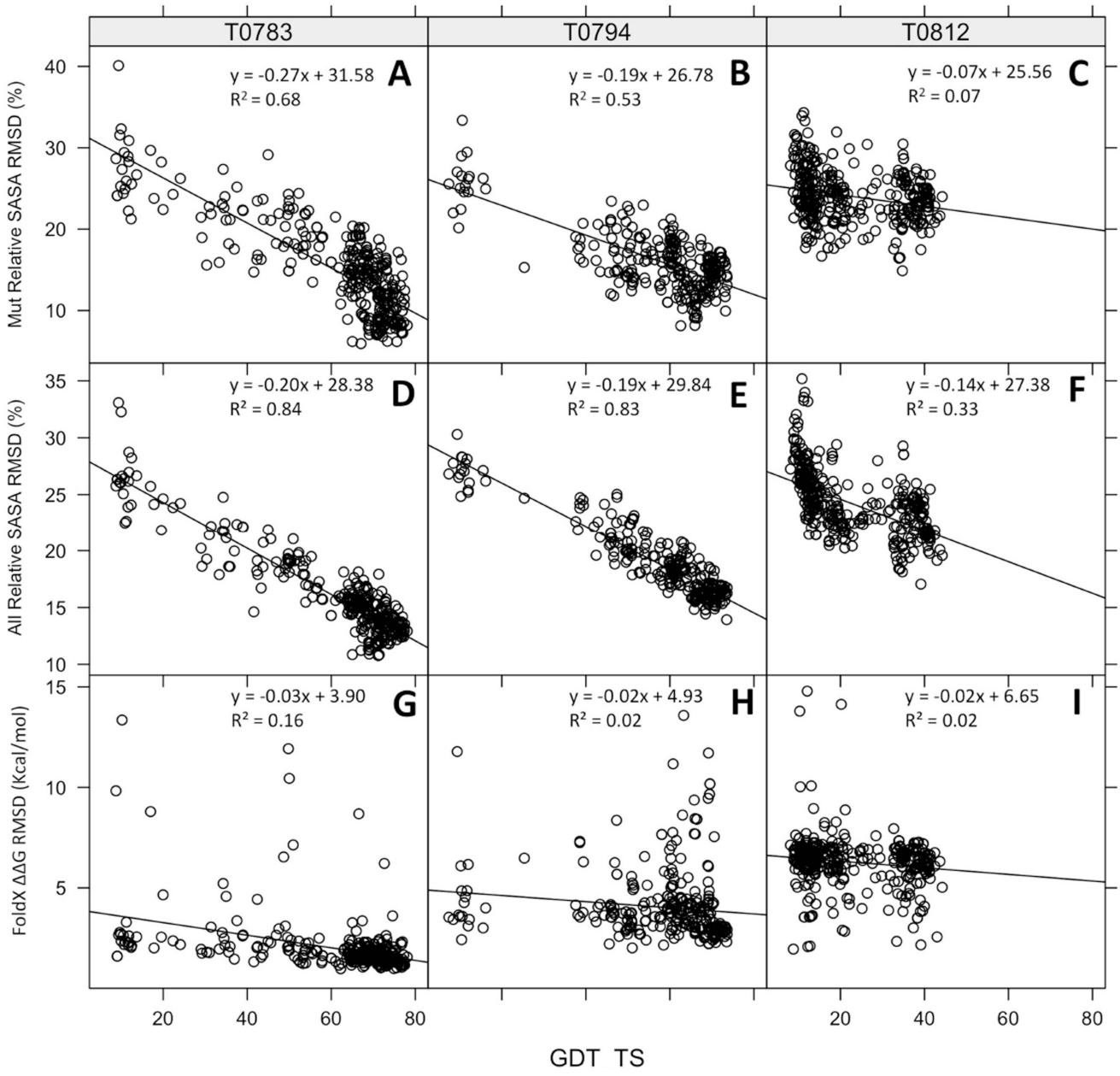

Figure 11.

Correlation of solvent accessible surface areas (SASA) and FoldX-derived ΔΔG of mutation with model accuracy for mutations in three human CASP11 targets. Correlation is measured by RMSD of SASA and FoldX values predicted on the models vs. those in the experimental structure. A. SASA RMSD vs GDT-TS for known mutation site residues in T0783 (Uniprot ISPD_HUMAN); B. SASA RMSD vs GDT-TS for known mutation site residues in T0794 (Uniprot VNN1_HUMAN); C. SASA RMSD vs GDT-TS for known mutation site residues in T0812 (Uniprot LAMA2_HUMAN); D. SASA RMSD vs GDT-TS for all residues in T0783; E. SASA RMSD vs GDT-TS for all residues in T0794; F. SASA RMSD vs GDT-TS

for all residues in T0812. G. ΔΔG RMSD vs GDT-TS for T0783; H. ΔΔG RMSD vs GDT-TS for T0794; I. ΔΔG RMSD vs GDT-TS for T0812.

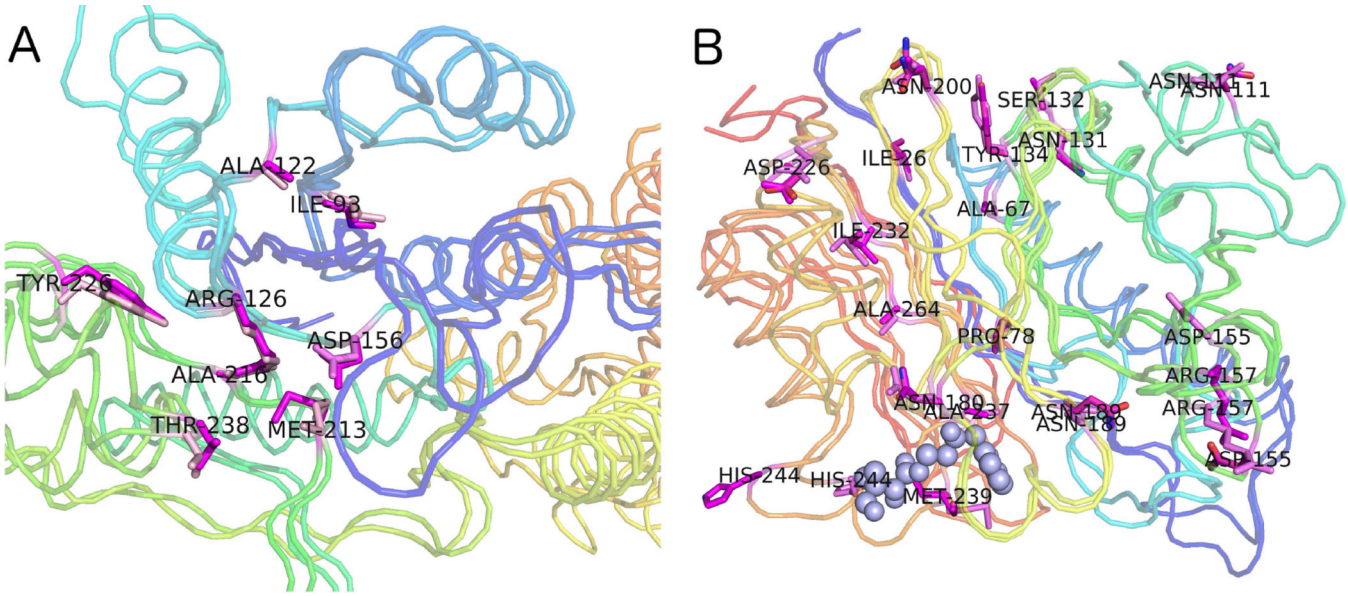

Figure 12.

A. Best model of T0783 (by group SEOK-refine) superimposed on experimental structure with mutations associated with Muscular Dystrophy-dystroglycanopathy (MDDGA7) shown in magenta sticks for the experimental structure and pink sticks for the model. The mutations are all clustered in the N-terminal domain adjacent to the active site. B. Best model of T0794 (by group LEER) superimposed on experimental structure with mutations in human VNN1 found in COSMIC and other sources (Table II) shown in magenta sticks for the experimental structure and pink sticks for the model. The effects of these mutations on the protein are unknown and are not necessarily associated with disease.

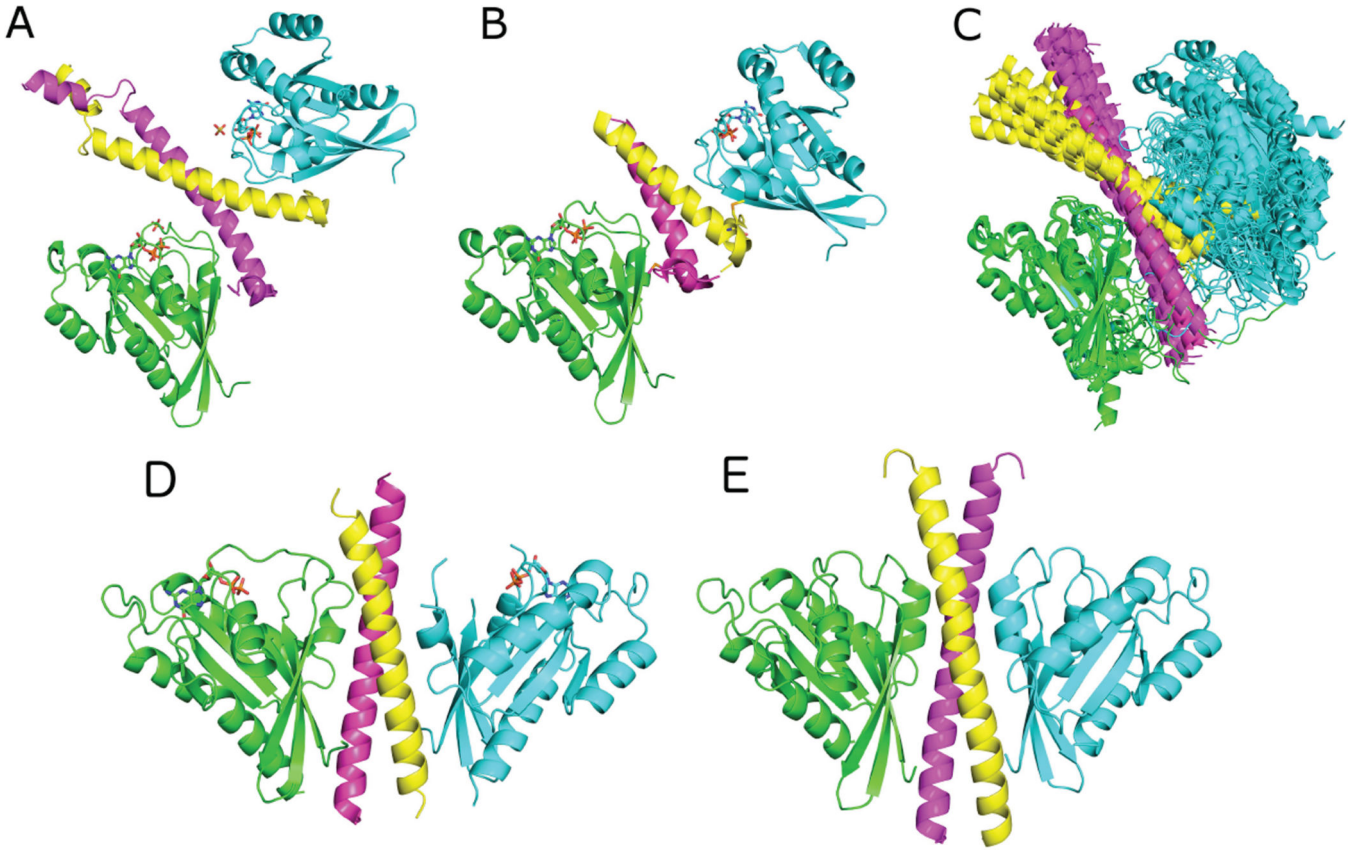

Figure 13.

Structures of Ras family proteins with coiled-coil effectors related to the heterotetramer target complex of CASP11 targets T0797 and T0798. A. Parallel, coiled-coil homodimer of Rab11 family-interacting protein 3 which interacts with two individual copies of Rab11a (PDB entry 2HV8); B. Parallel, coiled-coil homodimer of Rab11 family-interacting protein 2 which also interacts with two individual copies of Rab11a (PDB entry 2GZD). C. Most of the models produced by CASP predictors resemble these templates. D. CASP11 T0797/T0798 target heterotetramer (PDB entry 4OJF); E. Best model produced by predictors of T0797/T0798 (group Seok). In both images, the green monomer is oriented similarly to the green monomers in panel A.

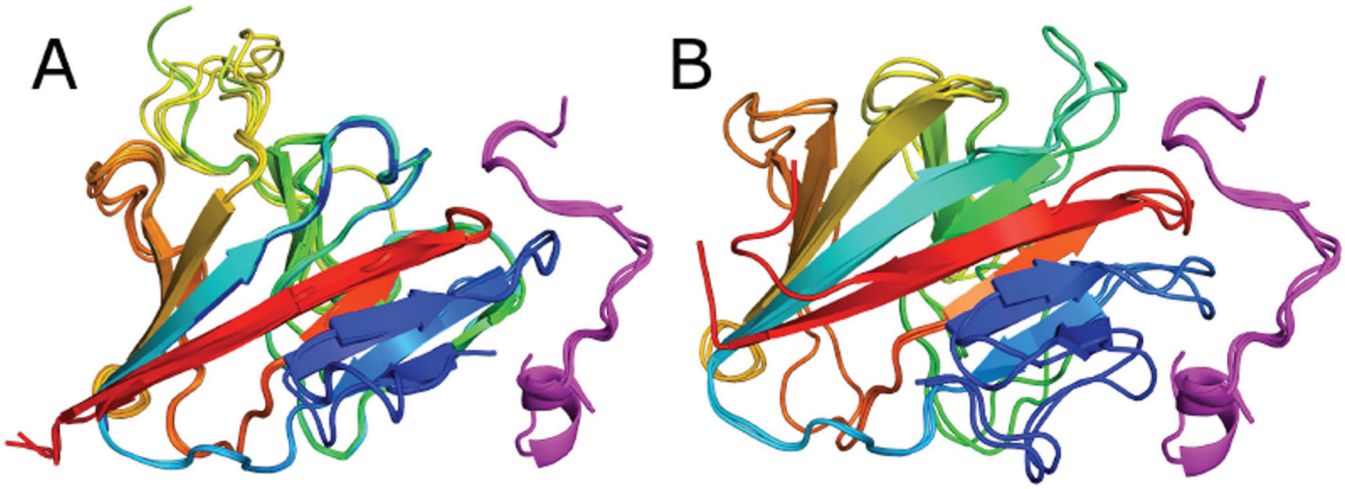

Figure 14.

A. ProtCID cluster of peptides bound to SPRY domains in the PDB; B. Experimental and predicted structure of CASP11 target T0856 shown with peptides from ProtCID cluster, after superposition of T0856 model and experimental structure onto closest structure in ProtCID cluster shown in (A).

Table I

Homodimer ClusPro docking results of CASP11 predicted monomer structures

Target	UniProt	PDB	ASA	GDT-TS Ave	GDT-TS Max	Top5 %	Top10 %	All %	Pfam	ProtCID cluster CFs	ProtCID entries	% First PDB BA	% PISA BA
T0801	RFFA_ECOLI	4PIW	1875	77.3	88.9	52.4	67.9	73.3	DegT_DnrJ_EryC1	22/23	44/45	91	100
T0776	A7ABD4_9PORP	4Q9A	943	75.9	83.9	4.8	10.2	17.7	Lipase_GDSL_2	3/39	10/58	70	70
T0843	Q0H2×1_9ACTO	4XAU	2583	73.9	80.1	29.1	35.2	38.5	DegT_DnrJ_EryC1	22/23	44/45	91	100
T0819	Q92R63_RHIME	4WBT	3463	73.6	84.8	18.8	24.9	33.1	Amino_tran_1_2	122/138	289/321	99	99
T0851	Q8KNF6_MICEC	4XRR	2672	65.1	80.4	63.7	70.9	76.5	UDPG_MGDP_N,dh,C	21/21	33/33	100	97
T0792	OSKA_DROME	5A49	665	61.3	78.4	3.7	4.8	5.3	OST-HTH	2/5	2/5	100	100
T0849	D0LNW1_HALO1	4W66	1938	56.3	64.5	3.8	7.0	12.9	(GST_N)_(GST_C)	80/82	223/225	97	95
T0764	A6LCA7_PARD8	4Q34	2442	67.1	78.5	7.7	12.1	22.5	*Abhydrolase_6*	(*17/163*)	(*46/406*)	(*78*)	(*48*)
T0852*	C7M590_CAPOD	4W9R	1163	65.3	75.2	0.6	1.1	7.4	*Esterase*	(*8/30*)	(*10/50*)	(*60*)	(*50*)
T0770	Q8A5I0_BACTN	4Q69	1119	58.1	68.6	0	0	0	*SusD-like_3*	(*1/18*)	(*3/21*)	(*0*)	(*0*)

ASA is the surface area of interaction: (2*ASA(monomer) - ASA(dimer))/2. Top5, Top10, and All refer to the percent of predicted monomers that had a predicted homodimer with a similarity to the native structure of Q>0.30. “ProtCID cluster CFs” provides the number of crystal forms (CFs) that contain the dimer of the biological assembly of the CASP11 target and the total number of CFs for that Pfam. “ProtCID entries” provides the number of entries that contain the dimer of the biological assembly of the CASP11 target and the total number of PDB entries for that Pfam. If the biological assembly dimer of the CASP target is not shared by other crystal forms in the PDB with the same Pfam, the information from ProtCID is for the largest cluster for that Pfam and is annotated in italic type and within parentheses (bottom of table). %First PDB BA and %PISA BA are the percent of entries in the ProtCID given in the previous column such that the PDB’s first biological assembly (First PDB BA) (whether author or software) or the PISA biological assembly (PISA BA) contains the dimer.

* Docking was performed with only the D1 domain of T0852

Table II

Mutations in three human CASP11 targets evaluated for accuracy of rSASA in the modeled structures

T0783, ISPD_HUMAN						T0794, VNN1_HUMAN						T0812, LAMA2_HUMAN					
Mutation	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**	**Mutation**	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**	**Mutation**	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**
V87I	COSMIC	Kidney	2.6	D	0.56	T26I	Biomuta	AML	24.5	N	0.00	W1185F	COSMIC	Lung	40.3	D	0.85
M102R	COSMIC	Ovary	17.6	D	0.95	A67E	COSMIC	Adrenal	0.0	D	1.00	E1191K	COSMIC	Skin,UT	50.3	D	0.90
S107I	Biomuta	AML	44.4	N	0.47	P78S	Biomuta	Lung	0.0	D	1.00	T1205M	COSMIC	Lung	34.6	D	0.99
H114Q	Biomuta	AML	1.9	D	0.94	N111D	COSMIC	Endomet	37.9	N	0.00	V1211A	COSMIC	Endomet	34.5	N	0.00
K115N	Biomuta	AML	50.4	N	0.10	N131S	COSMIC	Breast	13.8	D	1.00	A1219V	COSMIC	Stomach	0.0	D	0.69
I117T	Biomuta	AML	0.4	N	0.10	S132P	COSMIC	Colon	39.8	D	1.00	M1221V	COSMIC	Esophag	3.1	N	0.02
S118P	Biomuta	AML	19.5	D	0.91	Y134C	COSMIC	Breast	10.5	D	1.00	D1222N	COSMIC	Melanoma	46.7	N	0.43
L119R	Biomuta	AML	9.6	D	1.00	D155H	COSMIC	Lung	17.9	D	1.00	M1224I	COSMIC	Lung	0.0	N	0.00
A122G	Biomuta	AML	0.0	N	0.00	R157C	COSMIC	Colon	11.1	N	0.09	D1227Y	COSMIC	Colon	11.4	D	0.73
A122P	Uniprot	MDDGA7	0.0	N	0.00	R157H	Biomuta	Breast	11.1	N	0.04	E1231D	COSMIC	Skin	28.9	D	1.00
V124G	Biomuta	AML	43.2	N	0.07	N180K	COSMIC	Breast	11.9	D	0.63	Q1238S	COSMIC	Skin	4.5	D	1.00
R126H	Uniprot	MDDGA7	21.2	D	1.00	N189K	Biomuta	Orophar.	13.9	D	0.89	O1240H	COSMIC	Esophag	10.6	D	1.00
A136V	Biomuta	Breast	9.8	D	0.53	N189S	Biomuta	Breast	13.9	D	0.53	K1252E	COSMIC	Skin	21.5	D	0.99
D156N	Uniprot	MDDGA7	5.1	D	1.00	N200H	COSMIC	Lung	49.1	D	0.77	I1257T	COSMIC	Endomet	2.1	D	0.97
E176K	Biomuta	Endomet.	39.4	D	0.62	D226Y	COSMIC	Endomet	31.0	D	0.99	A1261V	COSMIC	Ovary	0.0	D	1.00
A180V	Biomuta	Colon	0.0	D	0.97	I232T	COSMIC	Orophar	0.0	D	0.98	E1263G	Biomuta	AML	36.2	D	0.77
A182D	COSMIC	Lung	0.2	D	1.00	A237D	COSMIC	Rectum	3.0	D	1.00	E1263K	COSMIC	Lung	36.2	D	0.77
V191I	Biomuta	Orophar.	34.7	N	0.00	M239I	COSMIC	Lung	27.4	D	0.91	F1267V	COSMIC	Colon	67.9	N	0.00
R205C	COSMIC	Colon	44.6	D	1.00	H244Y	COSMIC	Endomet	59.1	N	0.02	R1277P	Biomuta	Orophar	34.9	D	1.00
M213R	Uniprot	MDDGA7	6.7	D	1.00	A264E	COSMIC	Rectum	0.0	D	0.99	T1280K	COSMIC	Breast	33.0	D	0.80
Q215P	COSMIC	Kidney	0.2	D	1.00							R1285I	COSMIC	Lung	39.0	N	0.08
A216D	Uniprot	MDDGA7	0.0	D	1.00							R1289N	COSMIC	Lung	13.2	D	0.97
Y226H	Uniprot	MDDGA7	0.0	D	1.00							H1290N	COSMIC	Lung	29.4	N	0.46
E235A	Biomuta	Endomet.	39.0	D	0.99							E1305K	COSMIC	Skin	35.4	D	0.99
E235K	COSMIC	Rectum	39.0	D	0.99							E1308A	COSMIC	Endomet	9.5	D	1.00
T238I	Uniprot	MDDGA7	53.9	D	1.00							E1310K	COSMIC	Colon	44.1	D	0.95
Y266C	Biomuta	Lung	33.1	D	1.0							R1322S	COSMIC	Skin	42.2	D	1.00

T0783, ISPD_HUMAN						T0794, VNN1_HUMAN						T0812, LAMA2_HUMAN					
Mutation	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**	**Mutation**	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**	**Mutation**	**Source**	**Pheno**	**%ASA**	**PP2**	**sPP2**
												R1326Q	COSMIC	Endo,Skin	50.7	D	1.00
												F1329L	COSMIC	Endo,Colon	0.1	D	1.00
												D1331N	COSMIC	Skin	45.9	N	0.27
												Y1334D	Biomuta	Lung	42.7	D	1.00

ASA is the percent accessible surface area, relative to an exposed amino acid of the same type. PP2 provides the Polyphen2 prediction (neutral or deleterious) and sPP2 provides the Polyphen2 score (probability of being deleterious). Pheno provides the organ site for mutations found in tumors (whether known to be driver mutations or not) or specific disease associations (such as MDDGA7).