Towards the prediction of protein interaction partners using physical docking

Wass M.N[1,2], Fuentes G[1,6], Pons C[4,5], Pazos F.[3] and Valencia A[1].

1. Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, 28029 Madrid, Spain.
2. Structural Bioinformatics Group, Centre for Bioinformatics, Imperial College London, London, SW7 2AZ UK.
3. Computational Systems Biology Group. National Centre for Biotechnology (CNB-CSIC). c/ Darwin, 3. 28049 Madrid, Spain.
4. Life Sciences Department, Barcelona Supercomputing Center (BSC), c/ Jordi Girona 29, Barcelona 08034, Spain.
5. Computational Bioinformatics, National Institute of Bioinformatics (INB), C/ Jordi Girona 29, Barcelona 08034, Spain.
6. Present address: Bioinformatics Institute. 30 Biopolis Street, #07-01, Singapore 138671.

SUPPLEMENTARY INFORMATION

Table of Contents-

The Background Set

The Background set contains 922 monomeric proteins that are representative at the SCOP superfamily level. The Biological unit from the pdb was used to class proteins as monomeric. Monomeric proteins are selected to ensure that the proteins are in an unbound form when used for docking. Analysis of the proteins in the background set has been performed in comparison to the benchmark set. This analysis includes surface area of the proteins, known interactions, species and number of domains.

All of the proteins in the background set are full length chains (not isolated domains). Figure S1 shows the number of domains present in the background set. Approximately two thirds of the background set are single domain proteins. The remainder generally have two or three domains.

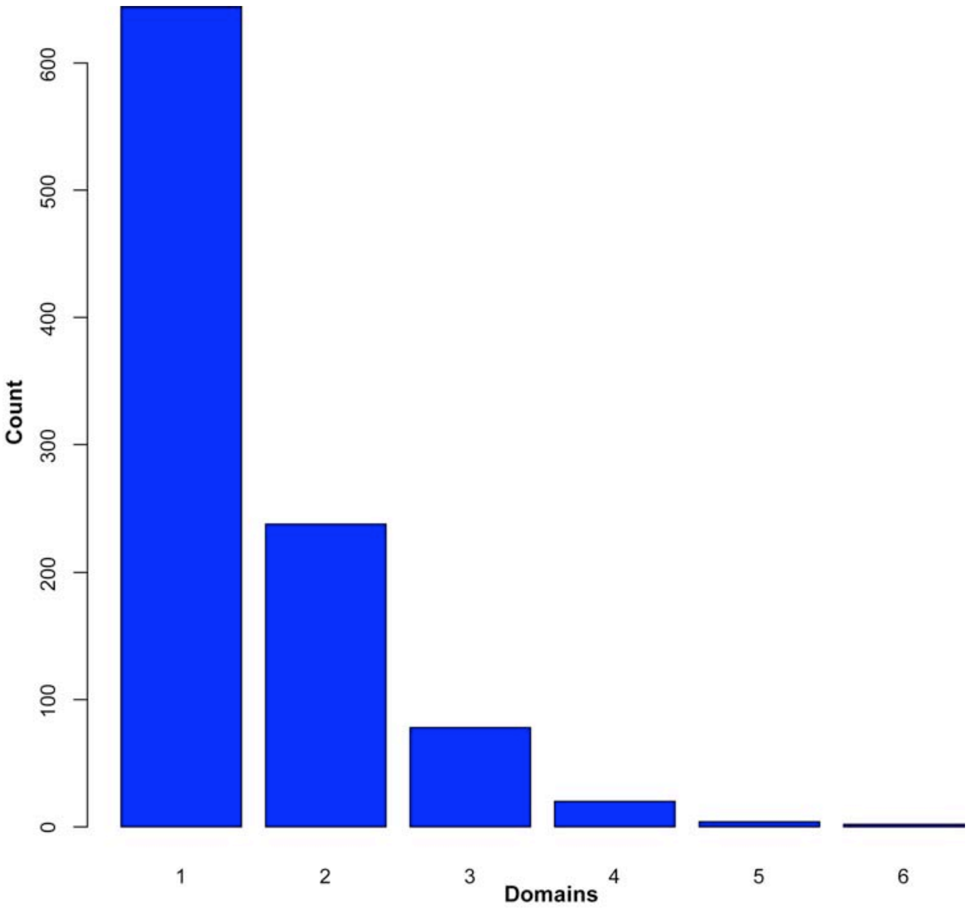


Figure S1. SCOP Domain composition of the background set.

Known interactions for the members of the background set were obtained from the Protein-Protein interaction databases MINT (Ceol *et al*, 2010) and IntAct (Aranda *et al*, 2010). Interactions were identified for 303 proteins, 291 of these had fewer than 70 interactions (see Figure S2). The remaining 12 had more than 70 interactions. Interactions for 35 of the 56 benchmark set proteins were also identified from the interaction databases (Figure S3). A similar pattern of interactions is present for proteins in the benchmark set (Figure S3), with most proteins having fewer than 10 interactions and very few with more than 30 interaction partners.

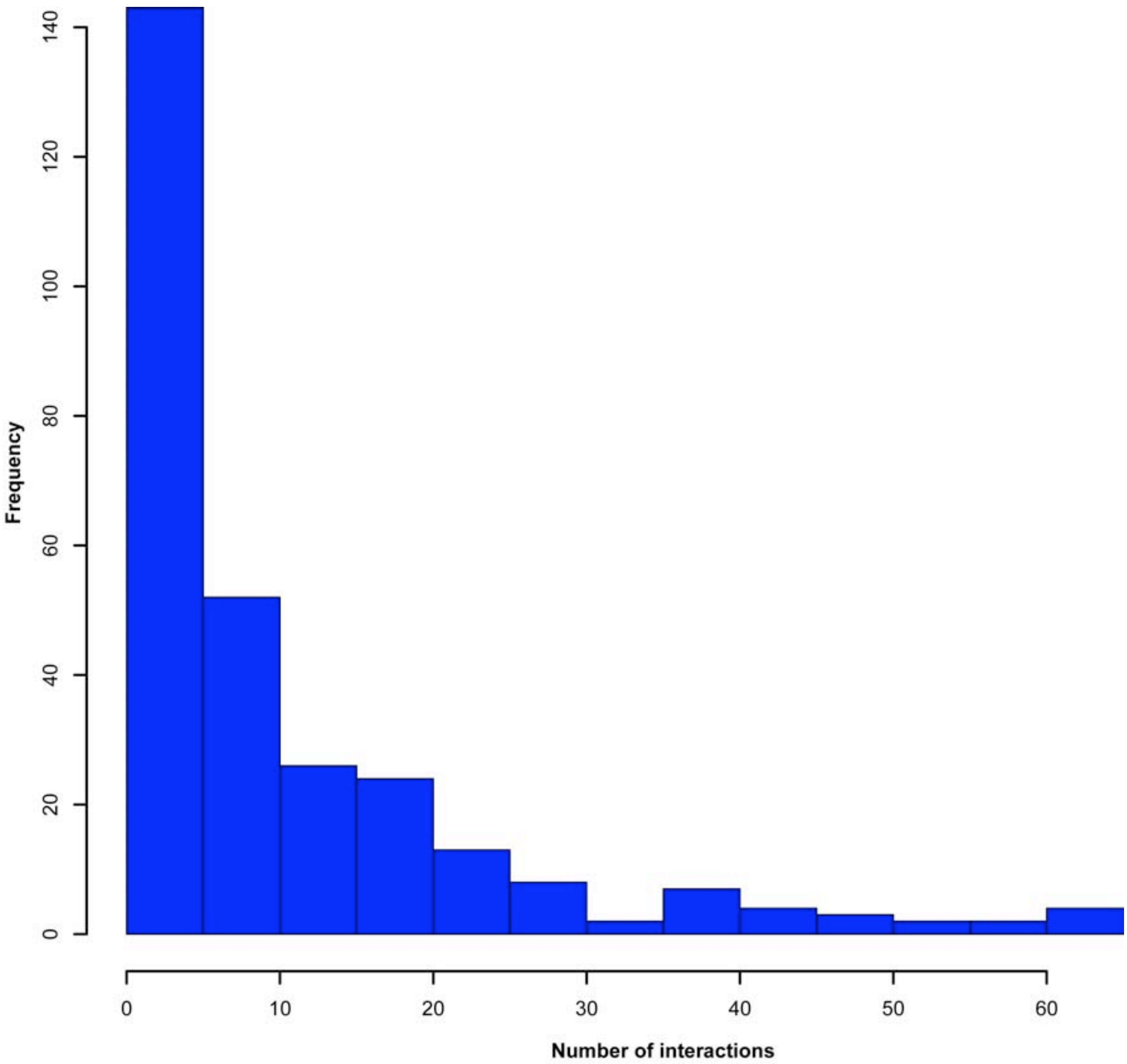


Figure S2. The number of interactions identified in IntAct and MINT for the background set proteins. Proteins with more than 70 interactions are not shown (12 in total).

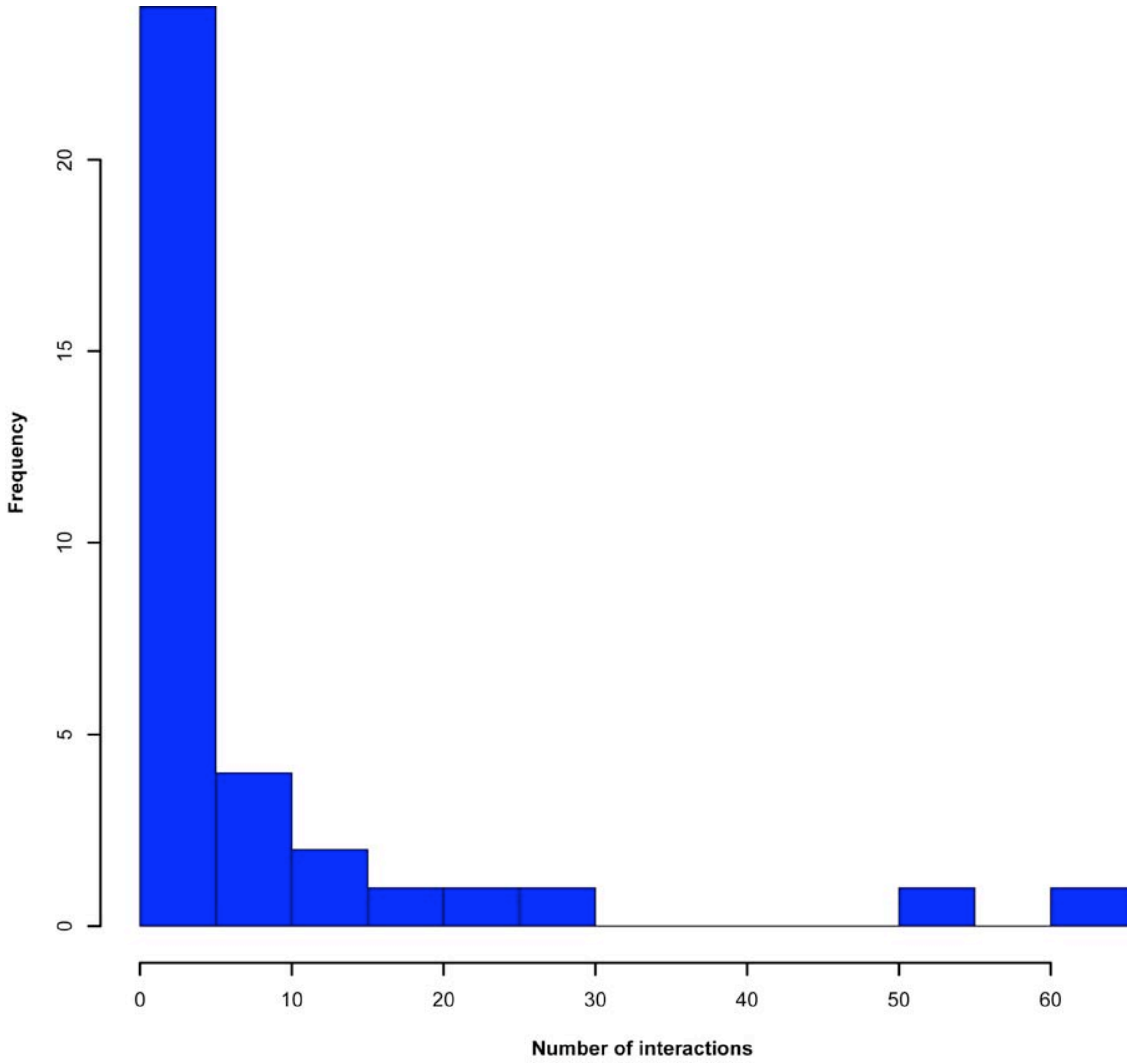


Figure S3. Interactions present in MINT and IntAct for proteins in the benchmark set.

The species that the background set proteins are from was also analysed and compared to the species of the proteins in the benchmark set. The 922 structures in the background set are obtained from 314 different species (Table SI). Structures from *Homo sapiens* are most widely present with 123 structures. *Homo sapiens* is also the source of 21 of the 56 benchmark structures (Table SII). The benchmark structures are from 21 different species (Table SII).

Taxonomy Id	Species	Number of structures
9606	Homo sapiens	123
562	Escherichia coli	93
4932	Saccharomyces cerevisiae	31
2336	Thermotoga maritima	26
274	Thermus thermophilus	24
10090	Mus musculus	20
10116	Rattus norvegicus	15
1423	Bacillus subtilis	15
300852	Thermus thermophilus HB8	13
1422	Geobacillus stearothermophilus	13
287	Pseudomonas aeruginosa	11
243274	Thermotoga maritima MSB8	11
9913	Bos taurus	10
602	Salmonella enterica subsp. enterica serovar Typhimurium	9
53953	Pyrococcus horikoshii	9
2234	Archaeoglobus fulgidus	9
7227	Drosophila melanogaster	9
3702	Arabidopsis thaliana	9
1773	Mycobacterium tuberculosis	9
2261	Pyrococcus furiosus	9
83333	Escherichia coli K-12	8
9031	Gallus gallus	7
727	Haemophilus influenzae	7
2190	Methanocaldococcus jannaschii	6
1280	Staphylococcus aureus	6
6239	Caenorhabditis elegans	5
469008	Escherichia coli BL21(DE3)	5
63363	Aquifex aeolicus	5
83332	Mycobacterium tuberculosis H37Rv	5
70601	Pyrococcus horikoshii OT3	5
176299	Agrobacterium tumefaciens str. C58	5
226186	Bacteroides thetaiotaomicron VPI-5482	4
10665	Enterobacteria phage T4	4
1396	Bacillus cereus	4
208964	Pseudomonas aeruginosa PAO1	4
242619	Porphyromonas gingivalis W83	4
1717	Corynebacterium diphtheriae	4
69014	Thermococcus kodakarensis KOD1	4
99287	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	4
9986	Oryctolagus cuniculus	4
139	Borrelia burgdorferi	3
8355	Xenopus laevis	3
226185	Enterococcus faecalis V583	3
210	Helicobacter pylori	3
83334	Escherichia coli O157:H7	3
10298	Human herpesvirus 1	3
196620	Staphylococcus aureus subsp. aureus MW2	3
4513	Hordeum vulgare	3
9823	Sus scrofa	3
446	Legionella pneumophila	3
271	Thermus aquaticus	3
5693	Trypanosoma cruzi	3
632	Yersinia pestis	3

Table SI. Taxonomy Analysis of the Background set. Only species that have three or more structures in the background set are displayed (54 species). Species that are present in the benchmark set are shaded grey.

Taxonomy Id	Species	Number of structures
9606	Homo sapiens	22
9913	Bos taurus	7
9823	Sus scrofa	5
4932	Saccharomyces cerevisiae	2
1390	Bacillus amyloliquefaciens	2
1280	Staphylococcus aureus	2
8618	Dendroaspis angusticeps	1
3847	Glycine max (soybean)	1
6421	Hirudo medicinalis	1
1168	Nostoc sp. PCC 7119	1
1932	Streptomyces tendae	1
10116	Rattus norvegicus	1
9940	Ovis aries	1
7067	Tenebrio molitor	1
287	Pseudomonas aeruginosa	1
4513	Hordeum vulgare	1
266	Paracoccus denitrificans	1
1887	Streptomyces albogriseolus	1
6253	Ascaris suum	1
1423	Bacillus subtilis	1
10299	Herpes simplex virus (type 1 / strain 17)	1
1582	Lactobacillus casei	1

Table SII. Taxonomy details of the benchmark set.

The UniProt subcellular location annotations of both the benchmark and background sets was considered to identify if there were similar proportions of intracellular and extracellular proteins in each. Sixteen of the 56 proteins in the benchmark set are secreted, compared to 72 of the background set. Therefore, although the proportion of extracellular proteins in the benchmark set is slightly higher, the majority of proteins in both sets of proteins are intracellular.

The surface area of the background and benchmark structures was also compared. Histograms of the accessible surface area calculated using DSSP (Kabsch and Sander, 1983) are shown in figure S4. The average surface area of the proteins from both sets is similar but the background set does have a few proteins that have larger surface areas than the proteins in the benchmark set.

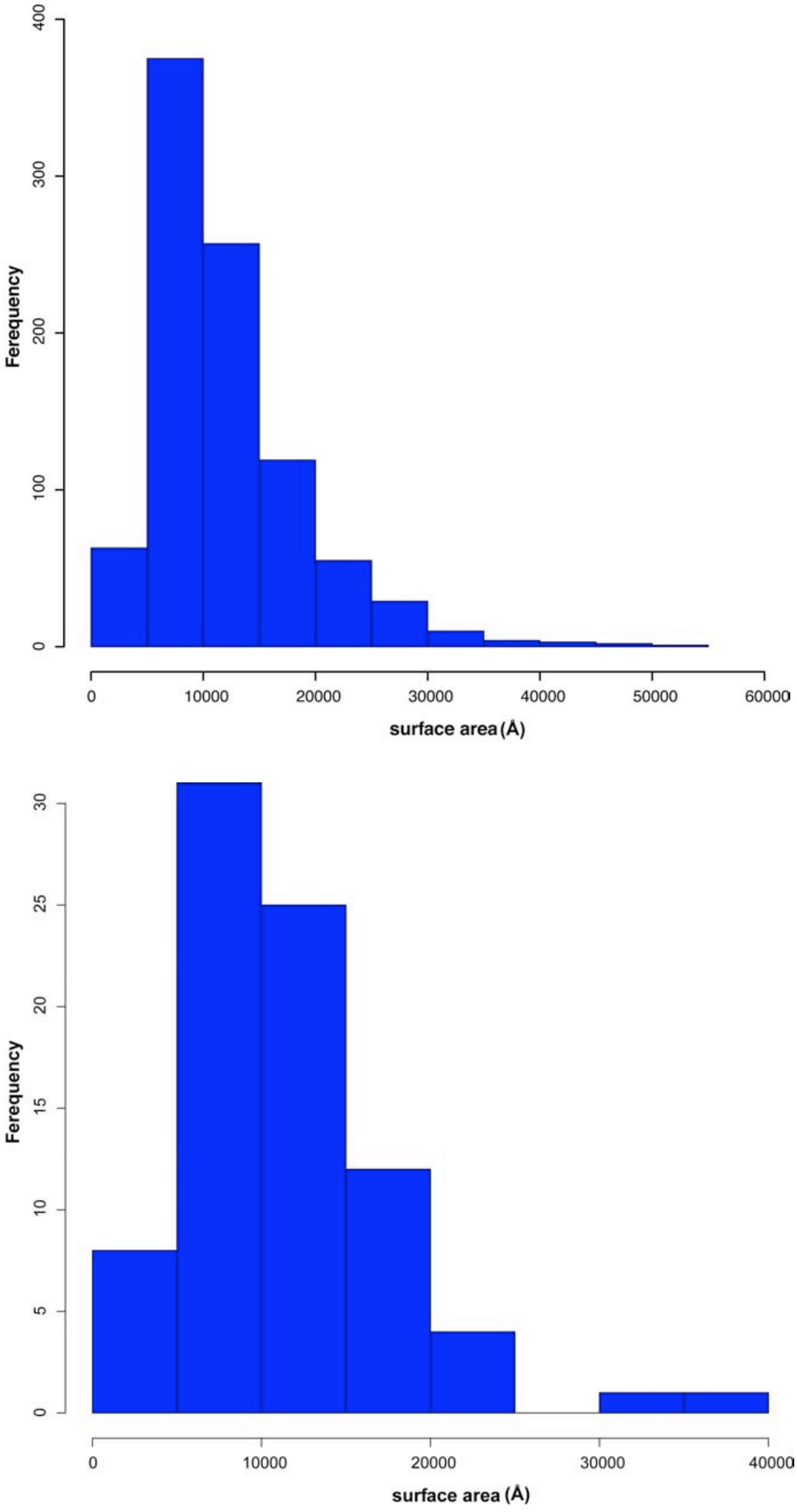
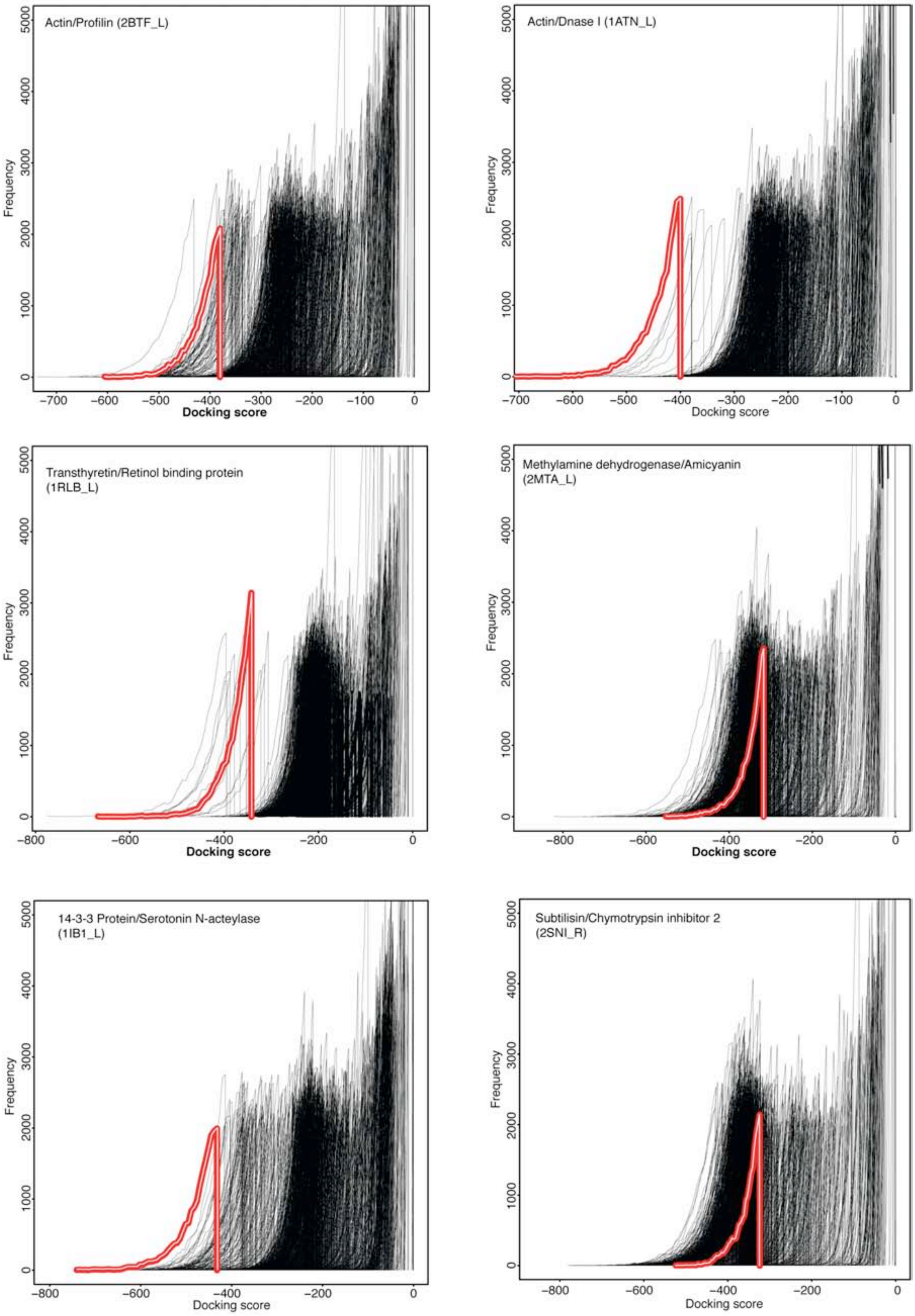


Figure S4. The accessible surface area of protein structures in A) Background structures B) benchmark structures. Note that the scales differ for the two histograms.

Results



Supplementary Figure S5. Benchmark and background docking score distributions. The docking score distribution of the benchmark complex (red) is plotted for Transthyretin/ Retinol binding protein (1RLB), Actin/Profilin (2BTF), (14-3-3 protein/Serotonin N-acteylase (1IB1), Subtilisin/Chymotrypsin inhibitor 2 (2SNI) and Methylamine dehydrogenase/Amicyanin (2MTA). In each graph, the docking score distributions of one of the components of the complex docked to background protein set are also plotted. The component used is indicated by the r (receptor) or l (ligand) after the benchmark complex id. For example 2SNI_r show the results for the docking of benchmark complex 2SNI (Subtilisin/Chymotrypsin inhibitor 2) and for subtilisin docked with the background set. The benchmark distributions are plotted in red and the background set in black. An image of the native benchmark complex is shown on each graph, the complex component being docked with the background set is colored red and the other complex component is blue.

Further analysis of results

For all of the docked benchmark complexes, poses have been identified that have high shape complementarity. This is shown in Figure S6, where the docking score distributions of the 42 benchmark complexes are plotted. The score distributions group together and with the exception of one of the complexes they have a score distributions within the range -800 to -300. This suggests that the docking program identifies high levels of shape complementarity between the proteins even for those examples where it does not result in accurate docking poses (close to the native interface) and even where few of the poses are in the general area of the binding site. This observation relates to our proposal that the results fit with the proposed funnel like intermolecular energy landscape in protein-protein interactions (McCammon, 1998) where proteins form non-specific encounter complexes before reorienting to their correct interface orientation (Blundell and Fernandez-Recio, 2006). The high levels of shape complementarity observed between the benchmark proteins may support this process of forming non-specific encounter complexes.

The docking distributions of the benchmark structures contrast with random pairs of proteins (i.e. the benchmark proteins docked with the background set) for which a much wider range of shape complementarity is observed from very low levels to those with greater shape complementarity than the benchmark complexes. Our results therefore show that the ability to distinguish the benchmark distributions from the background is not largely affected by the docking of the benchmark complexes but by the propensity for the benchmark proteins to have shape complementarity with the many members of the background set. Therefore it seems that some of the proteins in the benchmark set have surfaces that allow many proteins in the background set to have high shape complementarity and dock with them. It is possible that this may have biological relevance (i.e. these proteins may interact with many others). For example the majority of the proteases in the test set are not distinguishable from the background set (See protease section).

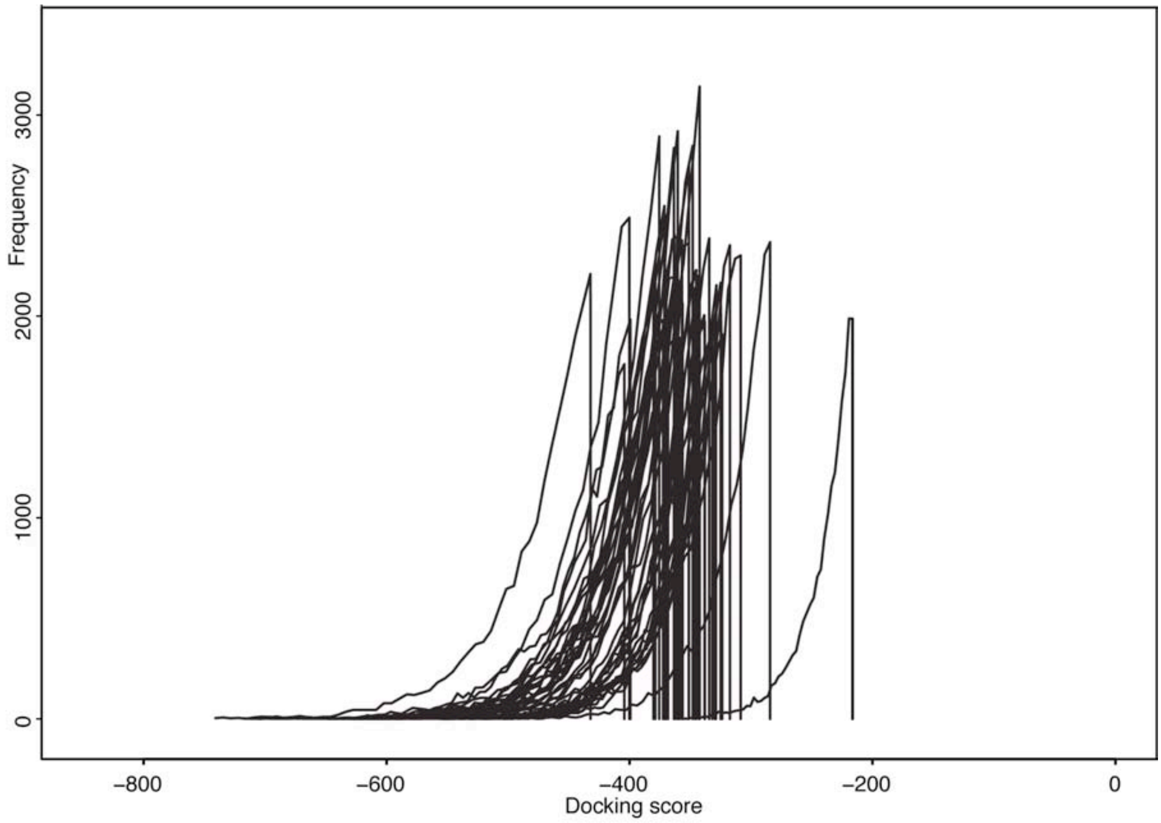


Figure S6. The docking score distributions of the 42 benchmark complexes.

The docking runs have been further analysed to identify if there are correlations between the results and the benchmark proteins and to investigate the use of different settings such as the number of docking poses used and the size of the background set.

The ability to distinguish the benchmark docking from the background set was compared to the size of the benchmark proteins (Figure S7). This graph shows that there is no correlation between the surface area of the benchmark protein and the percentage of background distributions that its docking distribution is better than.

The performance of the complexes against the background (meaning percentage of background set the benchmark distributions is better than) set was also compared to the RMSD between the bound and unbound forms of the benchmark proteins (RMSD data taken from the docking benchmark (Mintseris *et al*, 2005)). No correlation is observed between the conformational change of the complex and its performance against the background set (Figure S8). In fact some of the proteins with the largest conformational change on binding perform best.

For those benchmark complexes with known affinities (described by their dissociation equilibrium constants K_d (Kastritis and Bonvin, 2010)) their performance against the background set was plotted against their equilibrium constants (Figure S9). This shows that there is no correlation between the affinities of the interactors and how they perform compared to the background set.

Changes to the docking protocol were also considered. First the number of poses used to compare the benchmark and background distributions was changed. The analysis was performed using only 10,000, 5,000 or 1,000 poses compared to the 20,000 that were originally considered (Example distributions are shown in Figure S10). The results are invariant to these changes, with the same overall results as those for the original set (i.e. the results shown in Table 1). For individual complexes some minor changes were observed, with the total number of distributions that the benchmark complex was better than changing by one or two. These small changes are insignificant compared to the size of the background set and have no effect on the results.

For four of the complexes the number of poses retained was increased to 100,000. The docking score distributions are shown in Figure S10. The distributions show that there the results are effectively unchanged compared to those when fewer poses are retained.

The effect of the size of the background set on the results was also assessed. Random selections of 50, 100, 300 and 500 of the background set were chosen and the analysis performed on these subsets. Ten repetitions were performed, each using different sets of randomly selected proteins from the background set. The aim here is to identify how many structures are required in the background set to achieve stable results that do not fluctuate largely depending on the proteins selected to be in the background set. The results for each of the random sets was analysed in the same way as the original background set by calculating the percentage of the background set that the benchmark distribution is better than. The results are plotted in figure S11. In each graph in this figure the 10 results for each complex are displayed. We observe much less variation as the size of the background set is increased. This is partly because as the size of the background subset increases, it will have an effect on the percentage values (i.e. if the same number of distributions are better than the benchmark, this percentage will decrease with the increasing size of the background subset). What is of most interest here is to look at the variation in the results for individual complexes (i.e. vertical lines of points) at different sizes of background set to consider the effect this could have on the results of our analysis. With a background set size of 50, the results are highly sensitive to the choice of background proteins, this sensitivity reduces as the size of the background set is increased. For example if we were to choose a notional cutoff of 80% (say we predict interactions for any pair that achieves better than 80% of background), then with a background size of 50, 14 of the complexes have results from the 10 randomisations that are either side of this cutoff. In contrast only 4 complexes have results that are above and below this threshold when the size of the background set is 500 (Figure S11).

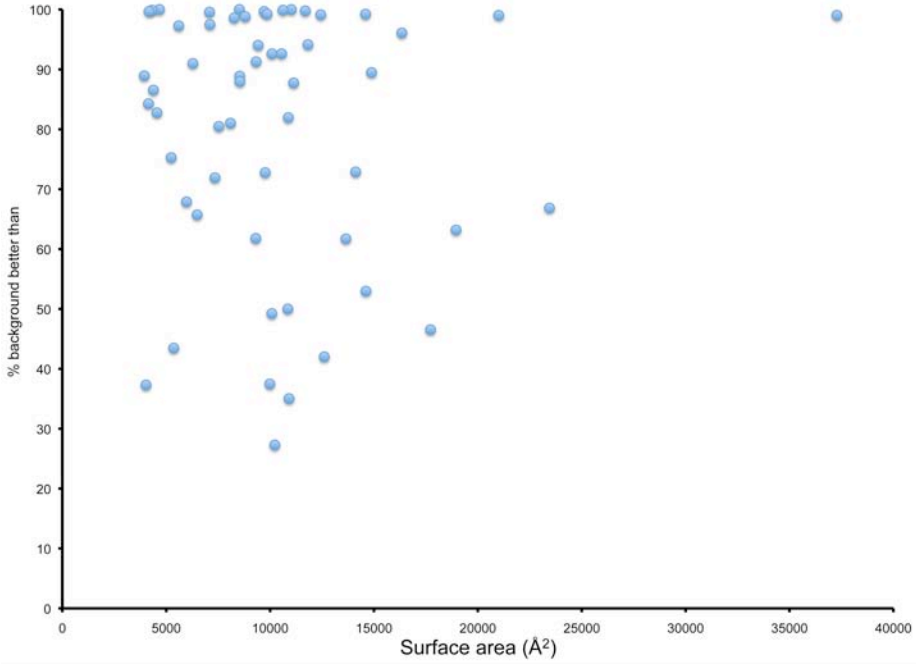


Figure S7. Performance against background set compared to benchmark protein surface area.

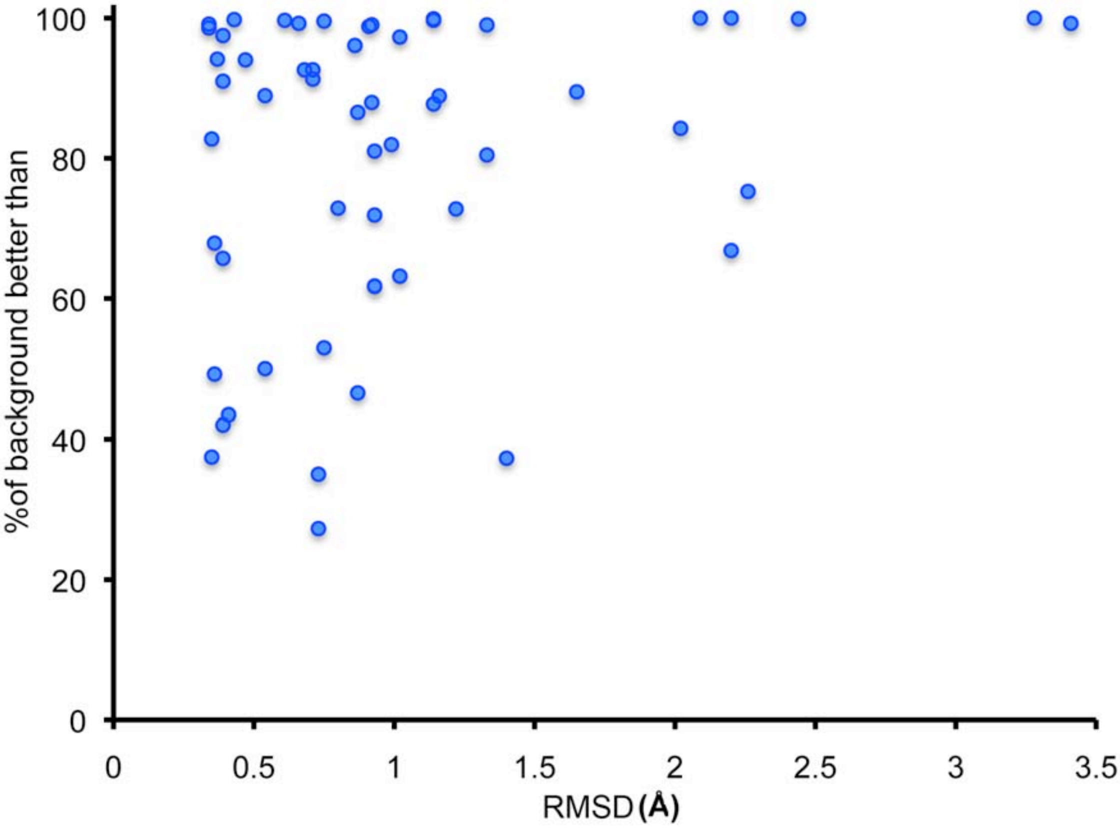


Figure S8. Comparison of performance against background set and the RMSD of the benchmark complexes. The percentage of the background set that the benchmark distriubiton is better than is plotted on the y axis and the RMSD between the bound and unbound conformations of the benchmark proteins on the x axis.

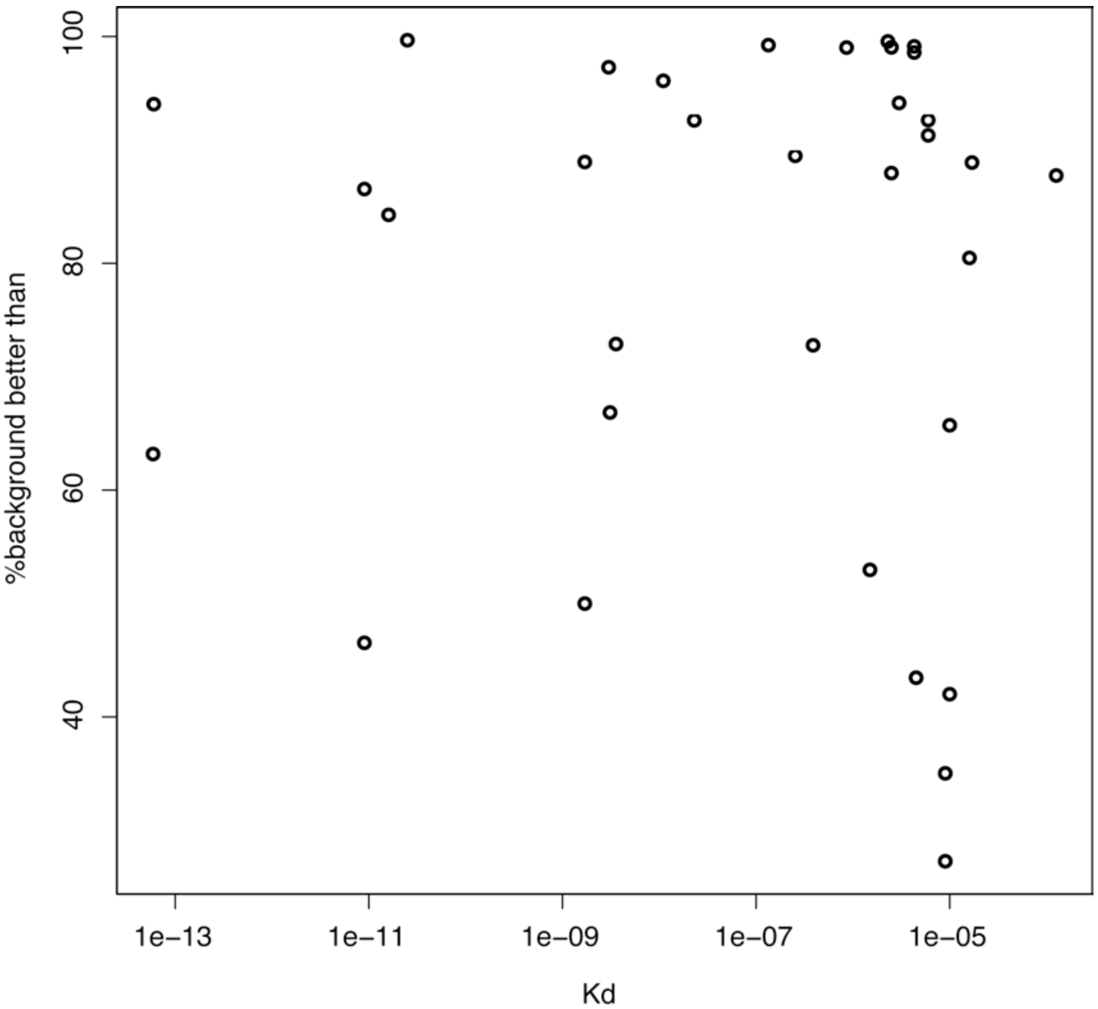
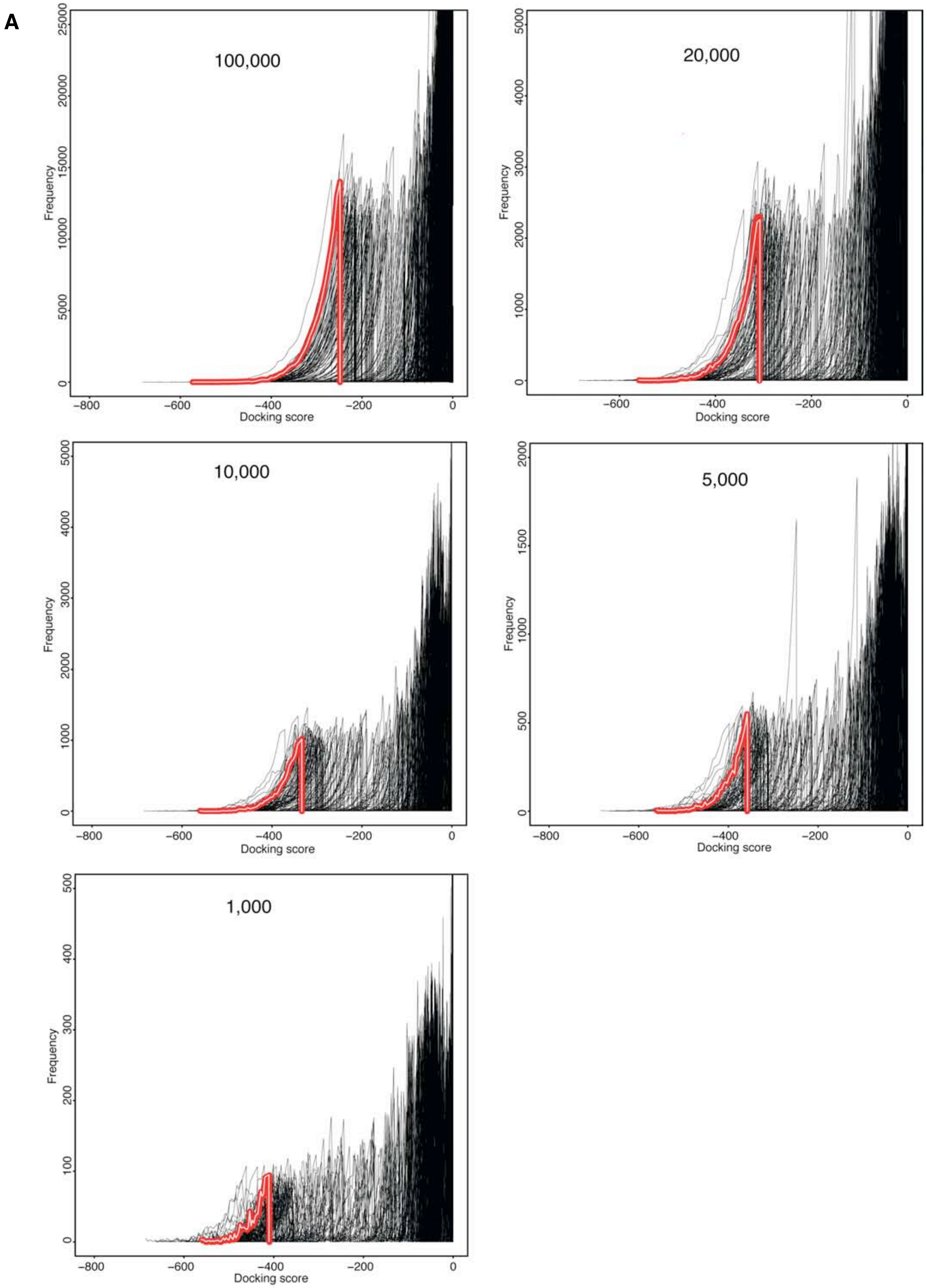
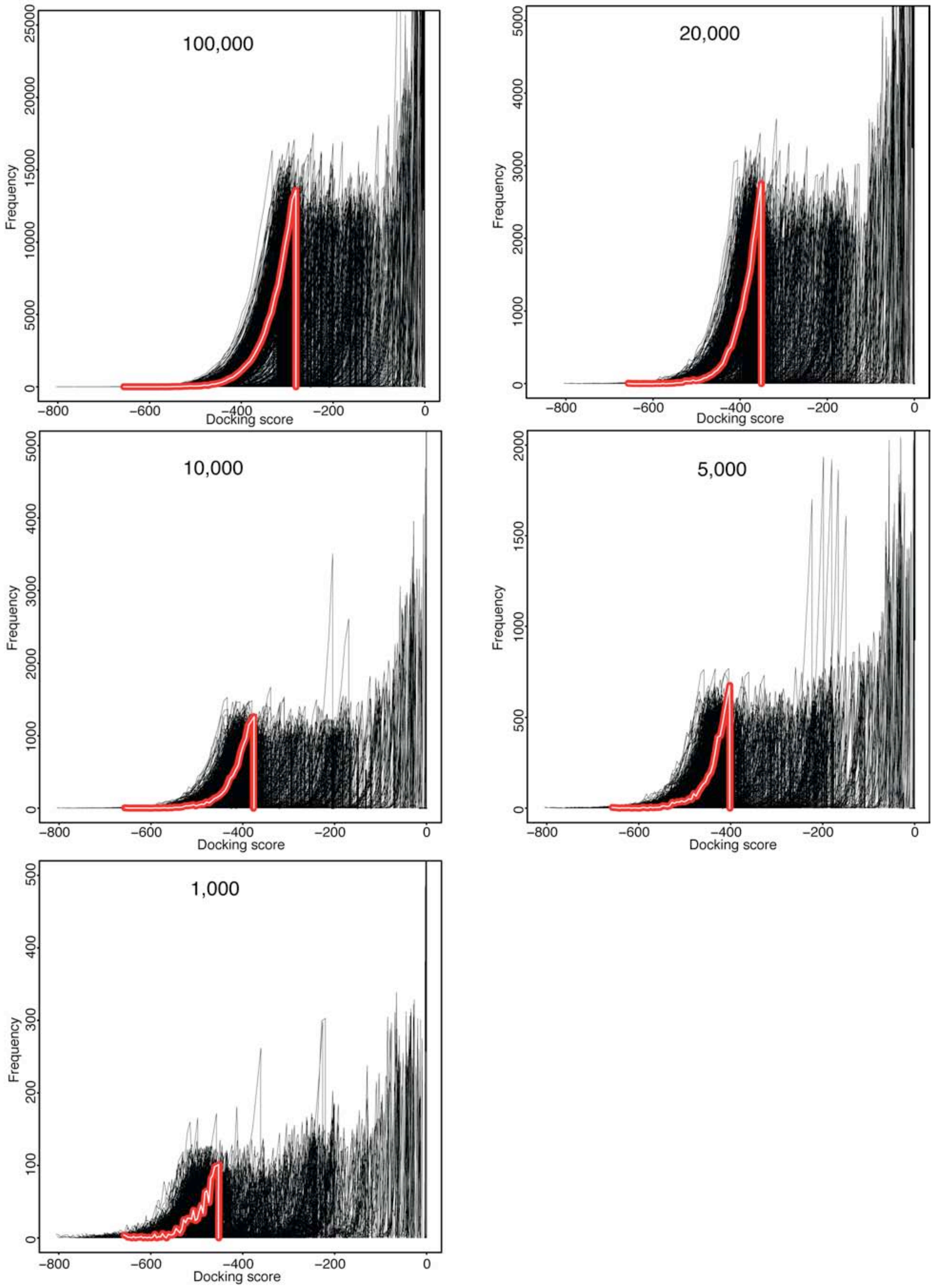


Figure S9. Comparison of benchmark complex performance and the equilibrium constant of the complex.


A
100,000
Frequency
0
5000
10000
15000
20000
25000
-800
-600
-400
-200
0
Docking score
20,000
Frequency
0
1000
2000
3000
4000
5000
-600
-400
-200
0
Docking score
10,000
Frequency
0
1000
2000
3000
4000
5000
-800
-600
-400
-200
0
Docking score
5,000
Frequency
0
500
1000
1500
2000
-800
-600
-400
-200
0
Docking score
1,000
Frequency
0
100
200
300
400
500
-800
-600
-400
-200
0
Docking score

B


100,000
Frequency
0
5000
10000
15000
20000
25000
−800
−600
−400
−200
0
Docking score
20,000
Frequency
0
1000
2000
3000
4000
5000
−800
−600
−400
−200
0
Docking score
10,000
Frequency
0
1000
2000
3000
4000
5000
−800
−600
−400
−200
0
Docking score
5,000
Frequency
0
500
1000
1500
2000
−800
−600
−400
−200
0
Docking score
1,000
Frequency
0
100
200
300
400
500
−800
−600
−400
−200
0
Docking score

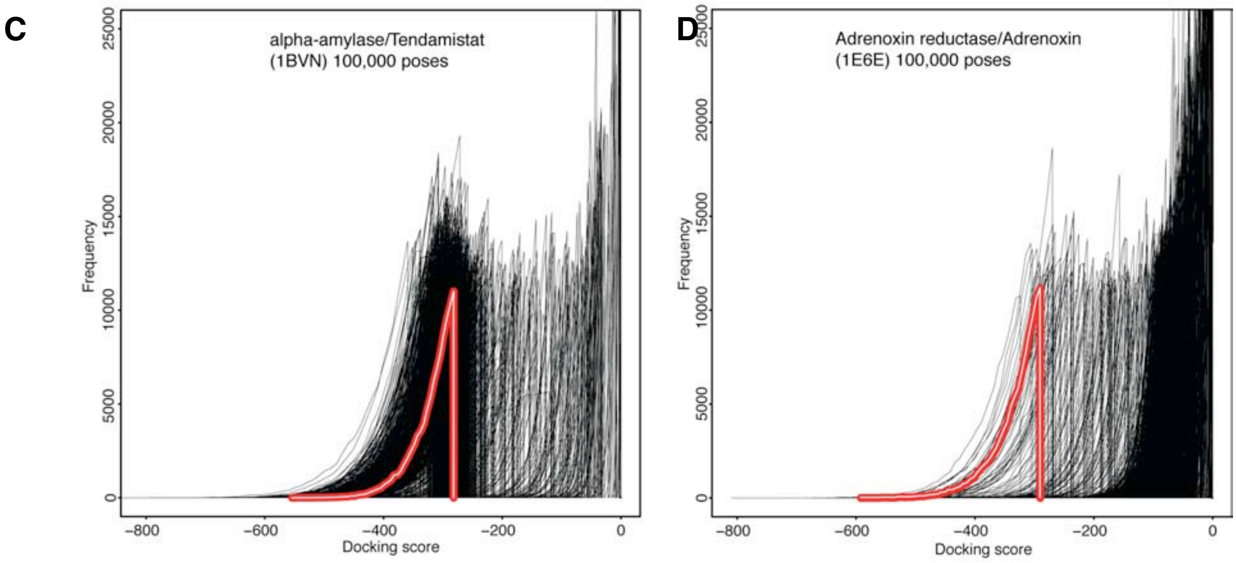


Figure S10. Docking score distributions using different numbers of poses. A) Docking score distributions for Ras GTPase/PIP3 Kinase complex (benchmark pdb code 1HE8) where 100,000, 20,000, 10,000, 5,000 and 1,000 top scoring models have been retained. B) As for A but for complex CDK2 kinase/Ckshs1 (benchmark pdb code 1BUH). C) Docking score distribution for alpha-amylase/Tendamistat complex where 100,000 poses have been retained. D) Docking score distribution for Adrenoxin reductase/Adrenoxin complex where 100,000 poses have been retained. Note that the number and locations of bins used to create the underlying histograms for each distribution is not set for the graphs where the number of poses retained is changed. Therefore the shape of distributions may change between each graph.

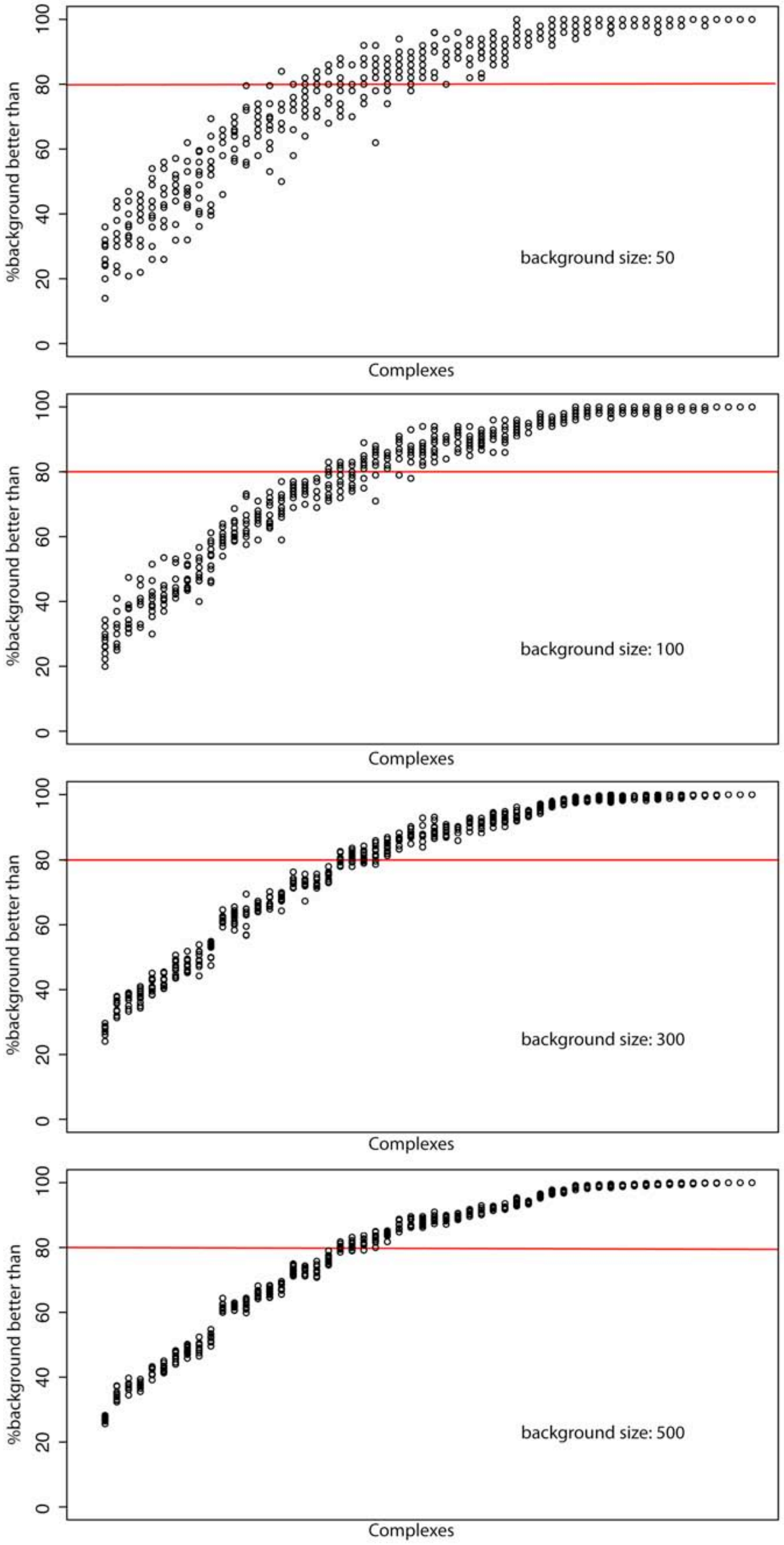
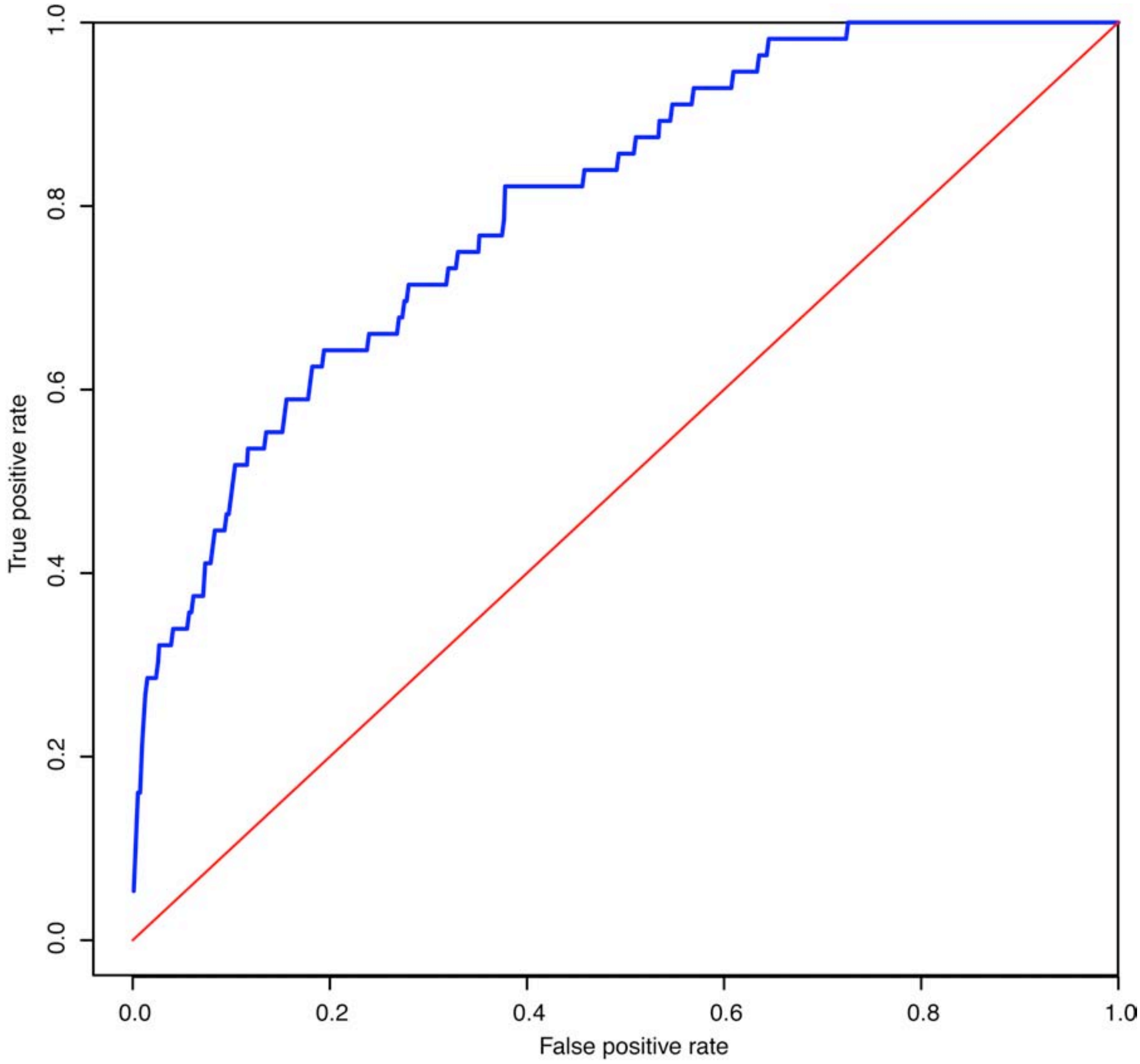
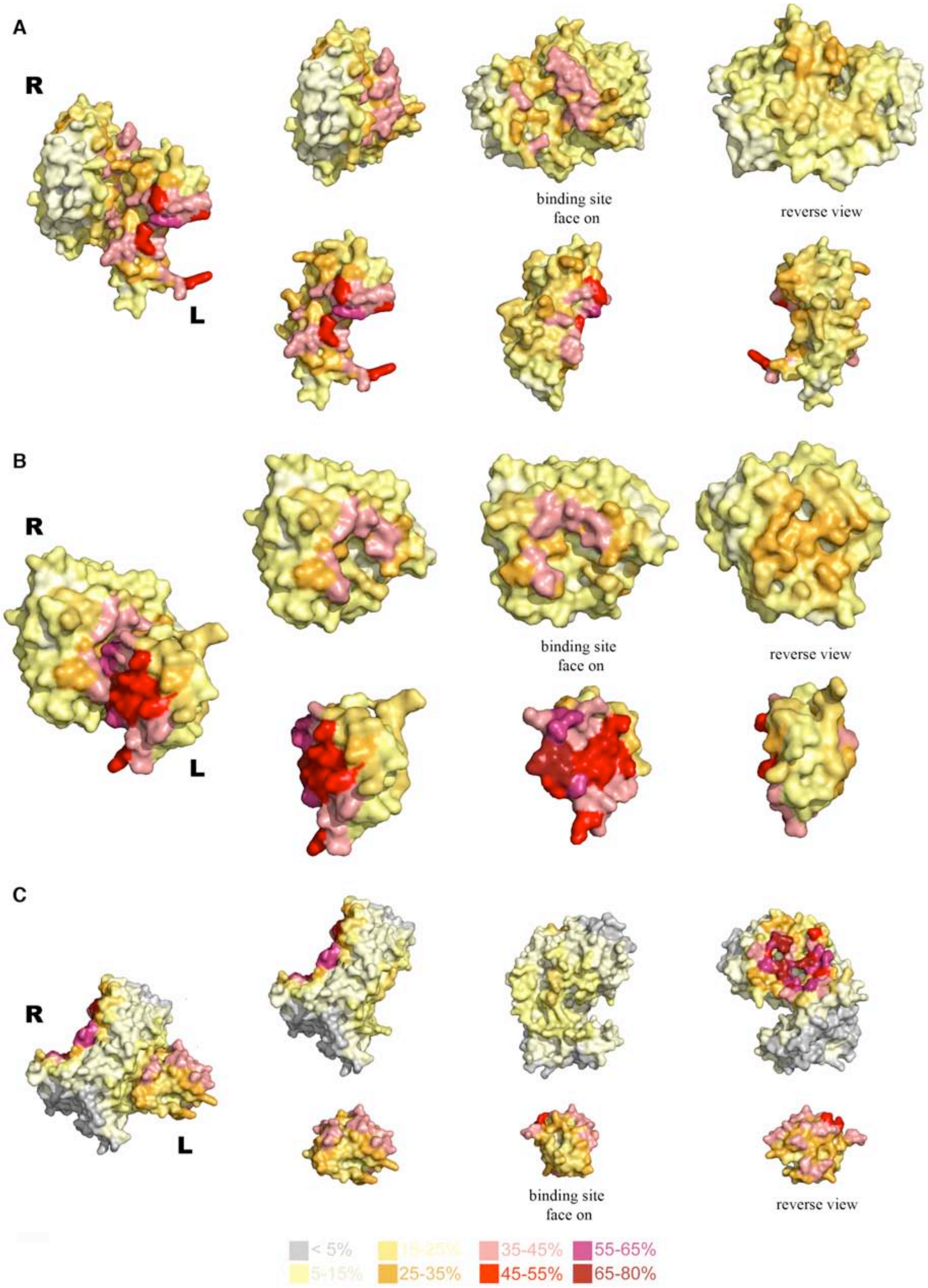


Figure S11. The effect of background set size on performance. The results obtained for the benchmark complexes using 4 sizes of background set. For each size, 10 random sets were selected and the analysis repeated. There are therefore 10 points per complex on each graph (displayed on a vertical line). The red line indicates a notional cutoff at 80% of the background set and illustrates the variance in the results between each of the different sizes of background set.

ROC Analysis



Supplementary Figure S12. ROC curve showing the relationship between the true positive and false positive rates of the method separating the 56 positives from the 51,632 negatives. The area under the curve (AUC) is 0.80. The diagonal represents a random method without discriminative power (AUC=0.50).



Supplementary Figure S13. Heat maps of the docking model putative binding sites. Heat maps indicating how often each residue is present in the binding site modelled by HEX were generated for each of the benchmark complexes. Example heat maps are shown for the following complexes (benchmark pdb id shown in brackets): A) Gt-Alpha/RGS9 (1FQJ), B) Subtilisin/Chymotrypsin inhibitor 2 (2SNI) and C) Adrenoxin reductase/Adrenoxin (1E6E). The unbound structures are shown and they have been aligned with their equivalent component in the native complex. The colour scheme as shown in the figure key indicates the percentage of HEX poses that a residue formed part of the putative interface.

Agreement of native binding site and docking patch	**Receptor**	**Ligand**
Binding site in largest patch	17 (40.4%)	14 (33.3%)
Binding site in 2nd patch	7 (16.7%)	4 (9.5%)
Partial overlap with binding site	1(2.4%)	11 (26.2%)
No overlap with binding site	17 (40.4%)	13 (31.0%)

Supplementary Table SIII. Agreement of putative binding sites with native binding site. The agreement between the patches of high intensity for the docked models with the native binding site is shown. Patches are defined as described in methods.

Protease Results

Most proteases (particularly Subtilisin) in the test set exhibit the second pattern of distributions. This may be due to the broad substrate specificity of such enzymes. It is possible that this is related to their non-specific hydrolytic function and the broad range of substrates that they target. This is the case of subtilisin (pdb entry: 2SNI), an extracellular alkaline serine protease that catalyses the hydrolysis of proteins with broad specificity for peptide bonds, and differs from the pancreatic enzymes by having a shallow binding groove on the surface, rather than the deep binding pocket of the latter (Perona and Craik, 1995). In principle, it is feasible that proteins in the background set may interact with it. It may therefore not be unreasonable that HEX has identified many good scoring complexes between these proteases and the structures in the background set.

Our assessment of the benchmark set makes a simple assumption that the native interactors from the docking benchmark do not interact with the proteins in the background set. For proteins that have few and or specific interaction partners, this assumption is likely to hold. However this assumption is more questionable for proteins that make many non-specific interactions as demonstrated by the proteases. It therefore seems reasonable that our method should be more likely to distinguish highly specific interactors and the presence of many background structures with good docking scores may be indicative of proteins that are involved in less specific interactions.

Using a Species-Specific Background Set

The full background set contains protein structures from 314 different species. Therefore, many of the proteins in this set are from different species to those in the benchmark set. In this analysis, species-specific background sets have been used for each of the benchmark structures. For each species the background set was generated in the same way as the full background with two differences – 1) the proteins were all from the same species, 2) to extend the number of structures that could be used, those without biounit information in the PDB but that only have a single chain present in the PDB were included.

The resulting species-specific background sets have varying sizes, the largest being that of *Homo sapiens*, which includes 339 structures. Only species for which more than 50 structures were identified were included, resulting in 32 of the 56 benchmark proteins being analysed. The size of the species-specific background sets is much smaller than the full background set and it is possible that the results could be affected by this.

As for the full background set the Wilcoxon rank sum test was used to compare the benchmark score distribution with those in the species-specific background set (Table SIV). These results were compared with the Wilcoxon rank sum test for the full background set, which shows that the results between the two background sets are very similar. If the difference in percentage of background set that the benchmark distribution is better than is considered (Table SIV), only 7 of the 32 structures have differences greater than 5%, with only 2 greater than a 10% difference. This demonstrates that the results are mainly consistent for the two background sets. For all of these 7 proteins the performance against the species-specific background set is better than against the full background set. In general the benchmark proteins perform slightly better against their own species set than the full background set, but this could be due to the size difference between the data sets.

This analysis shows that using a non species-specific background set does not have an appreciable effect upon the results observed and it is appropriate to compare the dockings of proteins from different species.

Benchmark Complex	Number of structures	number benchmark better than	%benchmark better than (species spec)	% benchmark (full background set)	Difference
1AK4_r	339	299	88.20	80.48	**7.72**
1AKJ_l	339	308	90.86	87.74	**3.11**
1ATN_l	67	67	100.00	100.00	**0.00**
1BUH_r	304	169	55.59	52.97	**2.62**
1CGI_l	339	305	89.97	84.27	**5.70**
1D6R_r	67	67	100.00	99.67	**0.33**
1E6E_r	67	65	97.01	99.02	**-2.01**
1E96_l	339	314	92.63	92.62	**0.00**
1E96_r	339	323	95.28	91.29	**3.99**
1EAW_l	67	60	89.55	88.94	**0.62**
1EAW_r	325	163	50.15	50.00	**0.15**
1FQ1_l	339	336	99.12	99.24	**-0.13**
1FQJ_l	67	66	98.51	98.81	**-0.30**
1GHQ_l	339	335	98.82	98.59	**0.23**
1GHQ_r	339	338	99.71	99.13	**0.57**
1GP2_r	85	79	92.94	89.48	**3.46**
1GRN_l	339	263	77.58	72.78	**4.80**
1HE1_r	339	234	69.03	61.79	**7.24**
1HE8_l	339	312	92.04	87.96	**4.07**
1HE8_r	319	317	99.37	99.04	**0.33**
1IJK_l	339	325	95.87	92.61	**3.26**
1KKL_l	79	79	100.00	99.57	**0.43**
1KTZ_l	339	315	92.92	91.00	**1.92**
1KTZ_r	339	338	99.71	97.51	**2.20**
1ML0_l	339	337	99.41	97.29	**2.12**
1QA9_l	337	116	34.42	27.28	**7.14**
1QA9_r	339	168	49.56	35.03	**14.52**
1RLB_l	67	67	100.00	99.24	**0.76**
1WQ1_l	339	309	91.15	88.89	**2.26**
2BTF_l	67	65	97.01	99.57	**-2.55**
2PCC_l	118	90	76.27	65.73	**10.54**
2PCC_r	116	60	51.72	42.01	**9.71**

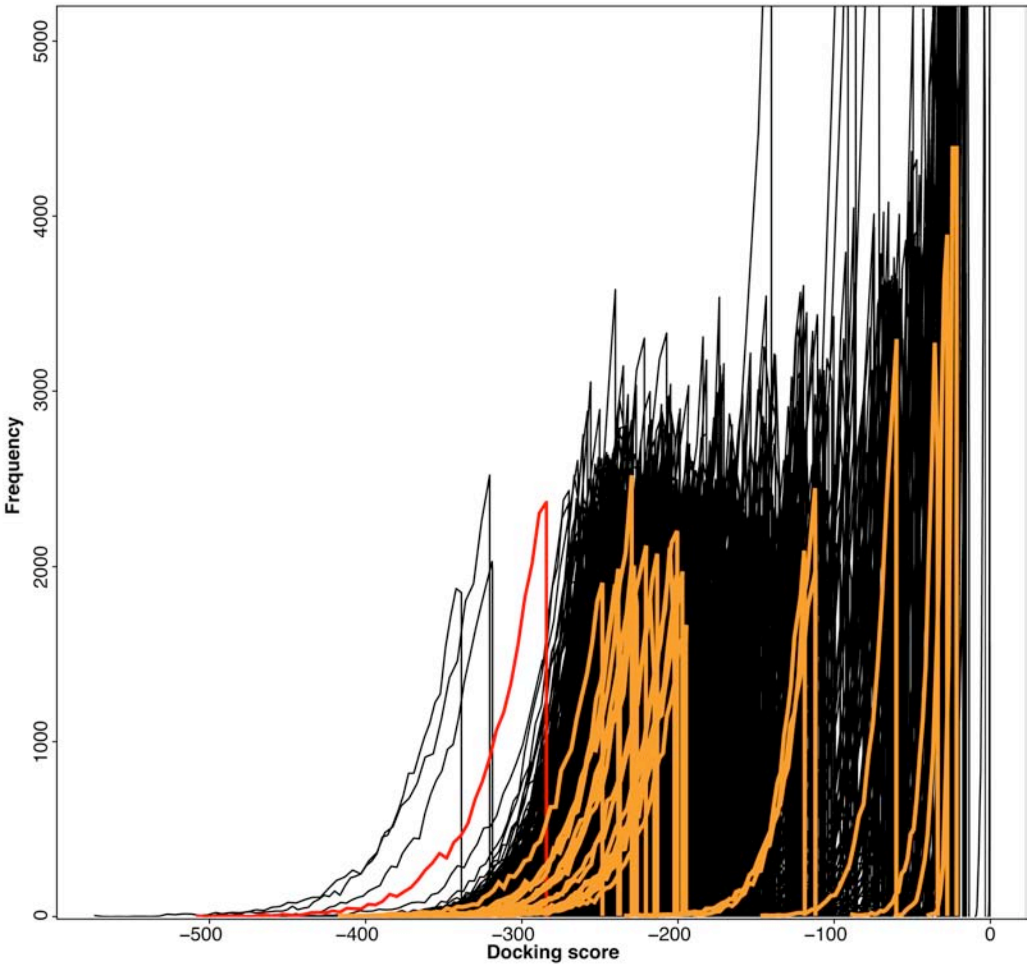
Table SIV. Wilcoxon rank sum test results for the species-specific background set. The wilcoxon rank sum test results are shown for the species-specific background set where there are more than 50 structures present for the species. The benchmark complex indicates the benchmark pdb complex that the benchmark structure originates from, followed by an r or l to indicate if the receptor or the ligand was docked with the background set. The table also displays the number of structures used for each, the number that the benchmark distributions was significantly better than (number benchmark better than), this value as a percentage and also the percentage of the full background set that the benchmark protein was better than. The difference column displays the difference in the percentage values for the species specific and the full background sets, with positive values indicating a higher percentage for the species-specific background set.

Docking for a single superfamily

Figures 3 and S15 show that for 3 examples the real interactors have a score distribution distinct from docking with structures from the same superfamily. Some of the score distributions from the same superfamily fall into the negligible score population, indicating that HEX has not been able to identify shape complementarity for some relatives of the known interactor. The majority of distributions fall into the medium score population, indicating that HEX has identified complementarity between them but not as strong as for the real interactors.

For the acetylcholinesterase/fasciculin complex DaliLite (Holm and Park, 2000) was used to assess the overall similarity between the structures used in the analysis and Acetylcholinesterase. When combined with the docking results (Fig. 3), it was surprising to find that the most structurally similar proteins (e.g. 1c7i an 1lpn) had some of the worst docking score distributions. In contrast some of those structures (e.g. 2axe, 1uxo and 1d07) most distantly related, in structural terms, to Acetylcholinesterase obtained the best score distributions. Visualisation of the structures gave some insight into this observation. Additional elements (such as α helices) within the area equivalent to the binding site in both 1c7i and 1lpn are likely to make these surfaces less favorable to docking with fasciculin. In contrast, the distantly related structures (i.e. 2axe, 1uxo and 1d07) show a more complementary morphology for the area equivalent to the binding site of the receptor, which may result in better docking scores. The ligand chosen to assess this docking experiment has a distinct morphological feature given by the tips protruding from its binding interface, thus shape complementary will be enhanced for proteins with two grooves matching this spatial restriction. The score distributions for this superfamily were also considered in the context of the background set (Fig. S14). None of the distributions from members of the superfamily are distinguishable from the background set with the exception of the native interactors from the benchmark complex. No correlation was observed between overall structural similarity and score distribution, this likely relates to differences in the morphology of the area equivalent to the receptor binding site, pointing to the feasibility of such an approach for the identification of putative interaction partners from amongst not only highly different proteins but also among structurally related proteins.

In the main text the Ras GTPase/PIP3 Kinase complex was used for analysis as the superfamily level. The Rac GTPase (present in a different benchmark complex – with p67 Phox) belongs to the same superfamily (P-loop containing nucleoside triphosphate hydrolases) as the Ras GTPase. Therefore it was possible to compare the docking of the Rac GTPase/p67 Phox complex at the superfamily level using the same structures within the superfamily were docked with p67Phox (Figure S15). Similar results as for the RasGTPase/PIP3 Kinase complex were obtained (Figure 3).



Supplementary Figure S14. Docking within a single superfamily. The docking score distributions of: the acetlycholinesterase and fasciculin benchmark complex (red), structures from the alpha/beta-Hydrolases superfamily and the fasciculin (orange) and fasciculin and the background set (black).

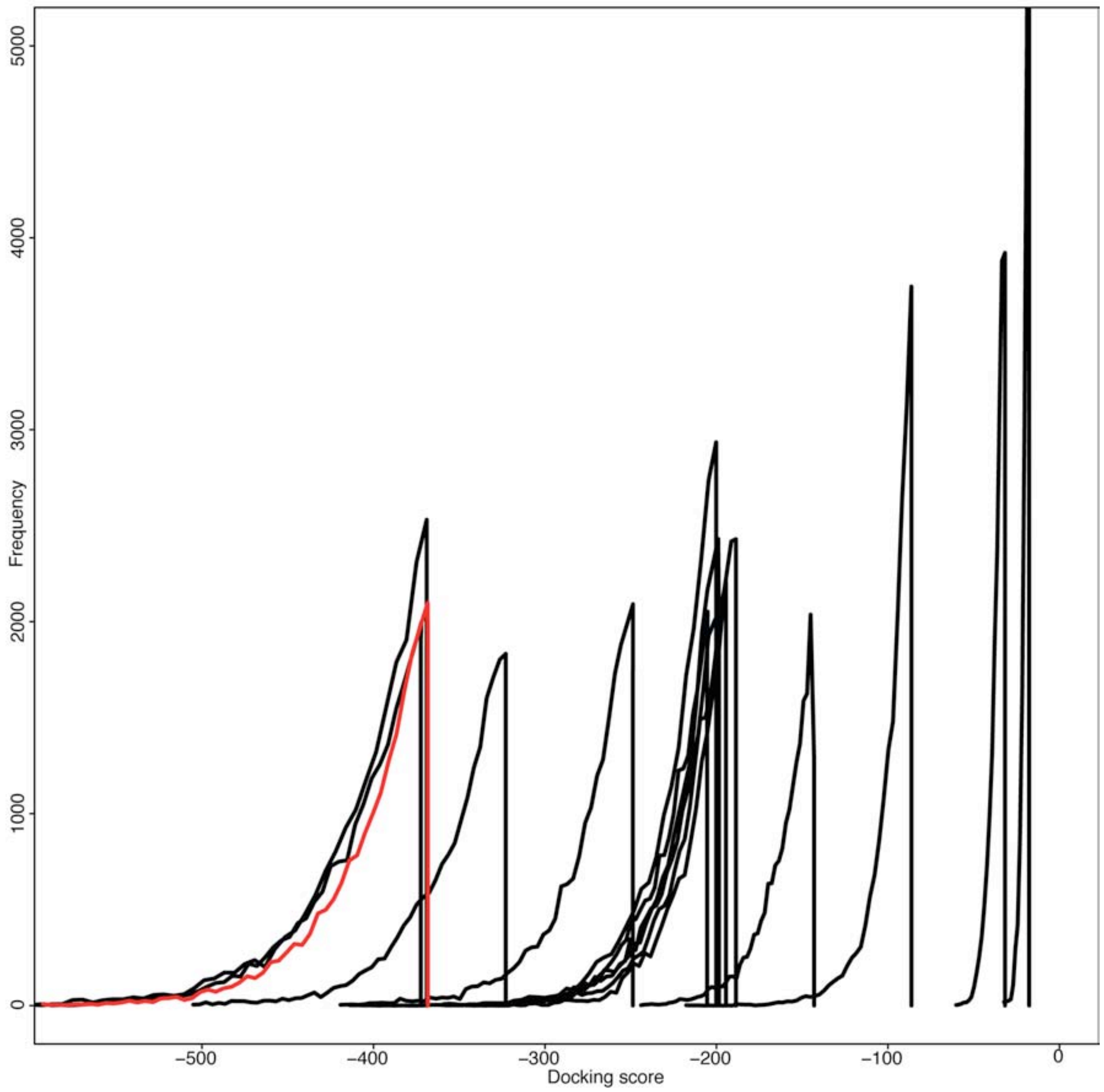


Figure S15. Docking score distributions for Rac GTPase/p67 Phox (red) and p67 Phox docked with other structures from the P-loop containing nucleoside triphosphate hydrolases superfamily (black).

Materials and Methods

Overview of Approach

An overview of our method is shown in Fig. S16. Individual proteins from the docking benchmark were selected. Each of these unbound structures was docked with its native interactor from the benchmark and additionally with a representative set of structures, representing most known protein superfamilies. These structures provide a random background of structures that are generally unlikely to interact with the proteins selected from the benchmark set. 20,000 poses were retained for the docking of the native interactors and for each of the background proteins docked with one of the interactors from the benchmark complex. Each model is associated with a pseudo-energy or docking score, which is a measure of how 'good' the model is. The numerous dockings are compared using the distributions of these scores, to assess if the score distribution of the known interactors is distinguishable from the random dockings generated for one of the components of the complex with a background set of proteins.

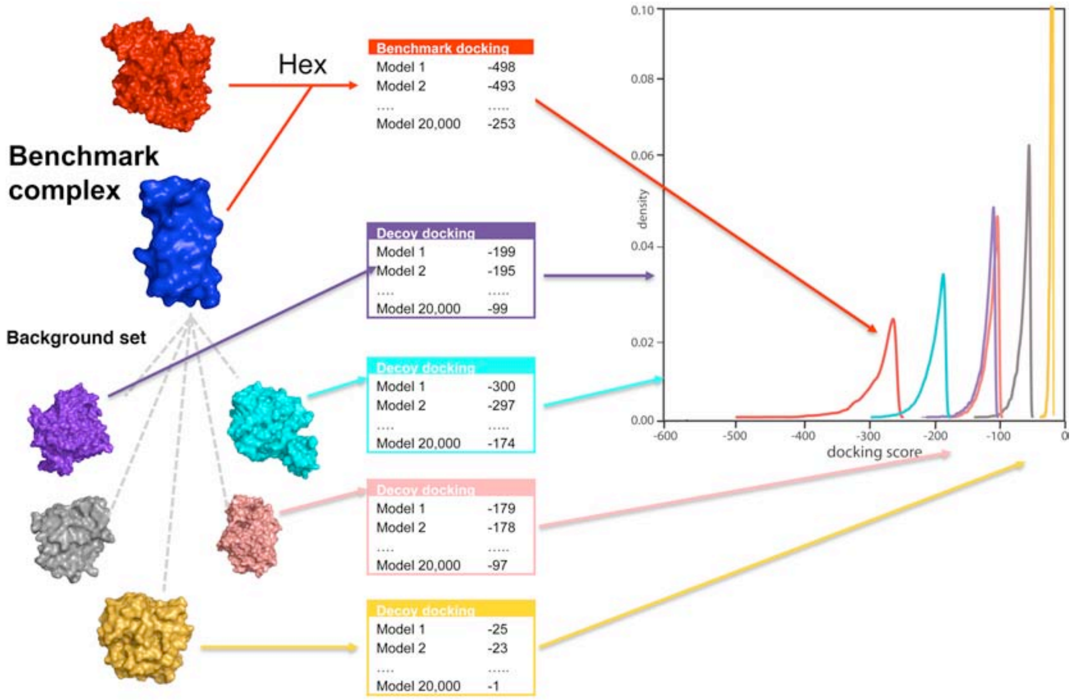
Benchmark Complexes

The following complexes from the docking benchmark were used in the analysis (pdb codes). The list indicates which of the complex components (receptor or ligand as defined by the docking benchmark) were used.

Receptor and Ligand: 1bvn, 1d6r, 1eaw, 1e96, 1f34, 1ghq, 1he1, 1he8, 1kkl, 1ktz, 1qa9, 2pcc, 2sic, 2sni.

Receptor only: 1ak4, 1avx, 1buh, 1e6e, 1eer, 1ewy, 1gp2, 1hia, 1tmq, 1udi.

Ligand only: 1acb, 1akj, 1atn, 1cgi, 1dfj, 1fq1, 1fqj, 1grn, 1ib1, 1ijk, 1klu, 1mah, 1ml0, 1rlb, 1sbb, 1wq1, 2btf, 2mta.



Supplementary Figure S16. Overview of High Throughput Docking Approach. Complexes from the docking benchmark are docked with each other. One of the unbound components of the complex (in this example the receptor) is additionally docked with a library of 922 different structures from SCOP (the background set). The distribution of docking scores obtained for the benchmark complex is compared to those obtained for docking with the background set.

Background data set

The background set of protein structures obtained from the pdb such that it represents SCOP superfamilies contained the structures in Table SV. See separate excel file. A brief legend is shown below.

Table SV. The Background set. The table lists all of the structures in the background set giving their pdb code, chain and identifier of the SCOP domain. PDB structures containing more than one SCOP superfamily have multiple entries.

Hex Docking Settings

Setting	Value
GRID_SIZE	0.6
DOCKING_RECEPTOR_SAMPLES	642
DOCKING_LIGAND_SAMPLES	642
DOCKING_ALPHA_SAMPLES	128
RECEPTOR_RANGE_ANGLE	180
LIGAND_RANGE_ANGLE	180
TWIST_RANGE_ANGLE	360
DOCKING_R12_RANGE	40
DOCKING_R12_STEP	1
DOCKING_R12_SUBSTEPS	2
DOCKING_MAIN_SCAN	16
DOCKING_MAIN_SEARCH	25

Table SVI. HEX docking settings.

Analysis of Interaction data

Known Interactions for the benchmark and background structures were identified from both MINT (release 5/5/2010) and IntAct (downloaded on 23/6/2010). The pdb codes were converted to uniprot accessions using the uniprot id mapping resource. Interactions for the uniprot accessions were then extracted from the interaction databases. The interactions identified from MINT and IntAct were then combined.

Calculation of accessible surface area

The accessible surface area of each of the proteins in the background and benchmark sets was calculated using DSSP (Kabsch *et al*, 1983).

Plotting docking score distributions

Figure S17 demonstrates how the docking score distributions are plotted. A histogram is first plotted and the distribution plotted based on this histogram by joining the x mid point of the bin with the frequency value for that bin.

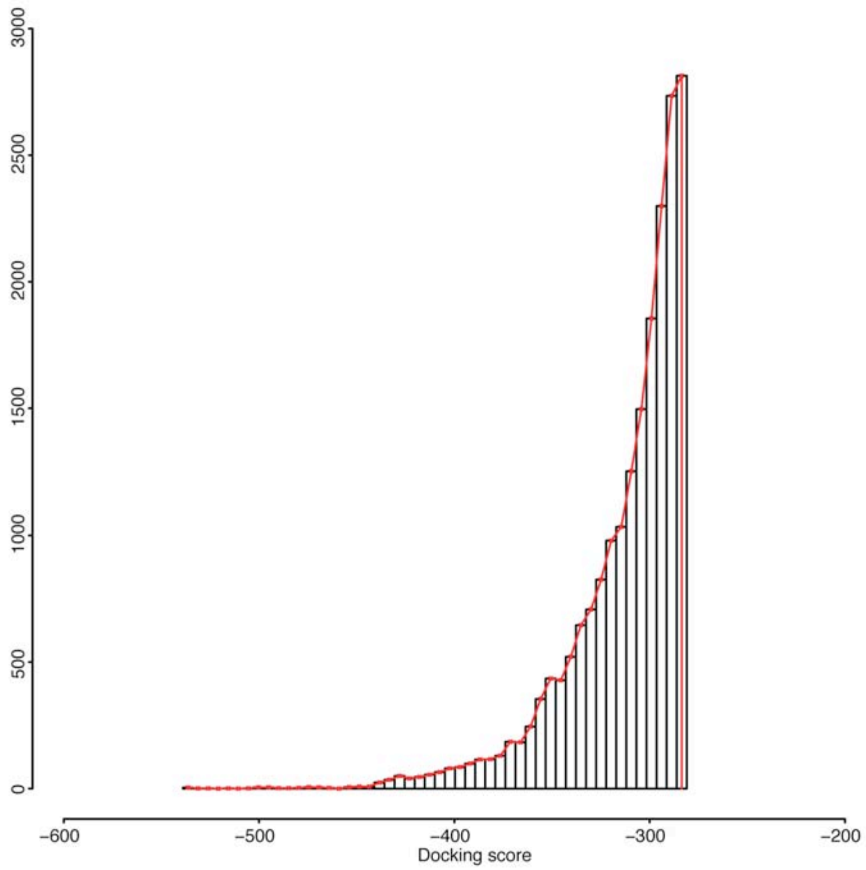


Figure S17. Plotting the docking score distributions. A histogram of the docking scores is first generated. The distribution is then plotted by placing a vertical line in the mid point of the highest value bin and then joining a lines between the mid points of each bin at the top of each bin as shown.

Supporting references

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H (2010) The IntAct molecular interaction database in 2010. *Nucl Acids Res* **38:** D525-531.

Blundell TL, Fernandez-Recio J (2006) Cell biology: brief encounters bolster contacts. *Nature* **444:** 279-280.

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the molecular interaction database: 2009 update. *Nucl Acids Res* **38:** D532-539.

Holm L, Park J (2000) DaliLite workbench for protein structure comparison. *Bioinformatics* **16:** 566-567.

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kastritis PL, Bonvin AMJJ (2010) Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *Journal of Proteome Research* **9:** 2216-2225.

McCammon JA (1998) Theory of biomolecular recognition. *Curr Opin Struct Biol* **8:** 245-249.

Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z (2005) Protein-Protein Docking Benchmark 2.0: an update. *Proteins* **60:** 214-216.

Perona JJ, Craik CS (1995) Structural basis of substrate specificity in the serine proteases. *Protein Sci* **4:** 337-360.