



Published in final edited form as:

J Chem Inf Model. 2016 June 27; 56(6): 1022–1031. doi:10.1021/acs.jcim.5b00387.

CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking and Scoring/Ranking Challenge

Richard D. Smith[‡], Kelly L. Damm-Ganamet[‡], James B. Dunbar Jr.[‡], Aqeel Ahmed[‡], Krishnapriya Chinnaswamy[¶], James E. Delproposto[¶], Ginger M. Kubish[¶], Christine E. Tinberg[§], Sagar D. Khare[§], Jiayi Dou[†], Lindsey Doyle[‡], Jeanne A. Stuckey[¶], David Baker[§], and Heather A. Carlson^{*,‡}

[‡]University of Michigan, Department of Medicinal Chemistry, 428 Church Street, Ann Arbor, MI 48109-1065

[¶]University of Michigan, Life Sciences Institute, 210 Washtenaw Ave, Ann Arbor, MI 48109-2216

[§]University of Washington, Department of Biochemistry, Seattle, WA 98195

[†]Department of Bioengineering, University of Washington, Seattle, WA 98195, Graduate Program in Biological Physics, Structure, and Design, University of Washington, Seattle, WA 98195

[‡]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. Seattle, WA 98109, USA

Abstract

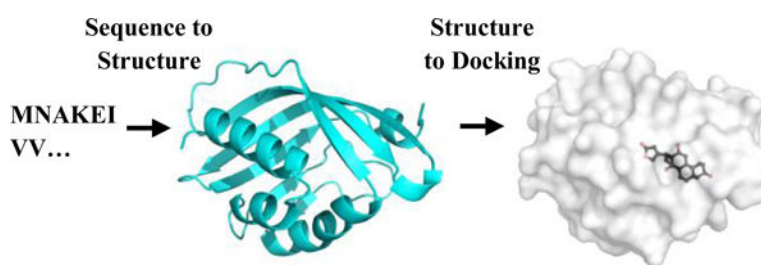
The Community Structure Activity Resource (CSAR) conducted a benchmark exercise to evaluate the current computational methods for protein design, ligand docking, and scoring/ranking. The exercise consisted of three phases. The first phase required the participants to identify and rank order which designed sequences were able to bind the small molecule digoxigenin. The second phase challenged the community to select a near-native pose of digoxigenin from a set of decoy poses for two of the designed proteins. The third phase investigated the ability of current methods to rank/score the binding affinity of 10 related steroids to one of the designed proteins. We found that eleven of thirteen groups were able to correctly select the sequence that bound digoxigenin, with most groups providing the correct three-dimensional structure for the backbone of the protein as well as all atoms of the active-site residues. Eleven of the fourteen groups were able to select the appropriate pose from a set of plausible decoy poses. The ability to predict absolute binding affinities is still a difficult task, as 8 of 14 groups were able to correlate scores to affinity (Pearson- $r > 0.7$) of the designed protein for congeneric steroids and only 5 of 14 groups were able to correlate the ranks of the 10 related ligands (Spearman- $\rho > 0.7$).

Graphical Abstract

*Corresponding Author: carlsonh@umich.edu.

Supporting Information Available:

The supporting information contains the experimental methods and protocols for protein expression, purification, and determination of the binding affinities by Thermofluor and Isothermal Calorimetry. The details of the crystallization of DIG5.1 (DIG18) are also provided. This material is available free of charge via the Internet at <http://pubs.acs.org>.



Introduction:

The main goal of many drug-design projects is to create a small molecule that uses a protein's innate ability for binding to achieve a desired physiological response.¹⁻³ These small molecules must bind to the target protein with a low equilibrium dissociation constant (high affinity).⁴ Over the past few decades, many computational methods have been developed that attempt to predict how well a small molecule will bind to a particular protein, improving time and cost efficiency in biochemical research.³ These methods continue to be honed and perfected, as not all methods function well for all protein systems. A large amount of structural and affinity information is necessary for continuing development, testing, and validation of these methods. In 2008, The Community Structure-Activity Resource (CSAR) center was organized with the goal of gathering a vast array of data for developing methods.⁵ This data is available to the community, both industry and academia, to help improve computational methods for drug-design purposes via the website www.csardock.org.⁶

The majority of drug-design projects have focused on new ligands based upon structural knowledge of a known protein target. A reverse objective, designing a protein to bind a particular small molecule, has also been of interest for creating protein-based biosensors, reagents, therapeutics, and diagnostics.⁷ Computational design of ligand-binding proteins is a challenging inverse test of docking and scoring methods, where experimental affinities can be fed back into the design process to refine the algorithms. Computational protein design and computational drug design share many similarities, and each requires a scoring function to rank-order a list of potential conformations to predict which is most biologically relevant and/or active.⁷ Over the past decade, the field of computational protein design has rapidly evolved, with successes ranging from the design of small hydrophobic cores to the creation of novel protein-protein interfaces to the construction of completely *de novo* proteins that perform specific functions.⁸ However, designing proteins to bind a small molecule has been a more challenging endeavor.⁹ Recently, three of the authors (C.E.T, S.D.K, and D.B.) developed a computational method in the framework of the Rosetta macromolecular modeling software¹⁰ to computationally design binding sites for the steroid digoxigenin.¹¹ Out of 17 designs that were tested experimentally, two proteins (DIG5 and DIG10) bound the target ligand digoxigenin with micromolar affinity. Affinity maturation by site-directed mutagenesis and screening using yeast surface display led to the identification of several variants (DIG10.1, DIG10.2, and DIG10.3) that were able to bind the ligand with affinities ranging down to sub-nanomolar levels. X-ray crystallography revealed the structures of the

two highest affinity binders, DIG10.2 (4J8T) and DIG10.3 (4J9A), bound to digoxigenin.¹¹ This work provided the unique datasets used for the 2013 CSAR Exercise.

Community-wide exercises have been utilized for a number of years to help the community develop and improve computational methods in areas such as protein-folding prediction (i.e. CASP¹²), prediction of protein interactions (CAPRI¹³), solubility prediction/solvation energies (i.e. SAMPL¹⁴), and structure-based drug design (i.e. the CSAR exercises^{15,16}, an evaluation by Warren *et al.* from GlaxoSmithKline¹⁷, the Teach-Discover-Treat initiative¹⁸, and the Kaggle Merck Molecular Activity Challenge¹⁹). Here, we used the data from the design and experimental characterization of digoxigenin binders¹¹, in conjunction with the CSAR blind-test framework, to run a community-wide exercise. This exercise was designed to challenge participants to predict which of the experimentally tested sequences would bind the small molecule digoxigenin (Phase 1), select a near-native pose of digoxigenin out of a set of docking decoys (Phase 2), and dock and rank-order/predict binding affinities for a set of ligands.

In this paper, the convention for labeling groups (A-V) is simply based on the order in which participants submitted answers. If a group submitted more than one set of results, usually from comparing multiple approaches, they were labeled as A-1, A-2, etc.

Methods and Materials:

Timeline:

The 2013 CSAR exercise was started March 25th, 2013 with the data available for download from www.csardock.org⁶. Participants were given approximately 1 month to complete this portion of the exercise, with submissions due by April 26, 2013 by uploading to the same website. Phase 2 of the exercise was started on May 13th, 2013 and submissions were due by May 31st, 2013. Phase 3 of the exercise started on June 21st, 2013 and ended August 30th 2013. Phase2 and Phase 3 data were provided and submitted using the same website as in Phase 1. Answers to each phase were given to participants at the conclusion of each phase. The entire exercise was completed prior to publication of the designed proteins and binding affinities of digoxigenin online in *Nature*, September 4, 2013¹¹. The X-ray crystal structures of DIG10.2 and DIG10.3 were released by the PDB²⁰ June 26, 2013, and DIG5.1 has been submitted, but has not yet been released at the time of this publication. The publication of the DIG10.2 structure was deposited during Phase 3 of the exercise; however, since the participants were given the coordinates of this structure as starting material for the phase, the release of the structure does not affect the results of this phase.

Phase 1:

Rules: Participants in the exercise were provided sixteen protein sequences that potentially bound the small molecule digoxigenin. The structure of digoxigenin is shown in Figure 1. They were asked to fold the protein sequence, dock digoxigenin, and predict which of the sixteen sequences had the ability to bind the ligand using methods of their choosing. Participants were also asked to send the resulting structure of the protein for comparison to the experimental structure.

Dataset Generation: Sixteen protein sequences (Table 1) were designed, constructed, and synthesized by C.E.T, S.D.K, and D.B. at the University of Washington. The ability of each of these proteins to bind digoxigenin was determined experimentally. Of the 15 designed sequences that did not bind digoxigenin, only 12 were used in phase 1. The other sequences were not included, as two of the sequences (DIG11 and DIG16) could not be expressed in *E. coli*, and one of the sequences (DIG15) showed binding in some assays that was not confirmed. In the dataset, we also included two protein sequences that were experimentally optimized to bind digoxigenin by directed evolution: DIG5.1 (labeled DIG18 in phase 1), which derives from the DIG5 design (PDBID: 5BVB), and DIG10.3 (labeled DIG19 in phase 1), which derives from the DIG10 design (PDBID: 4J9A)¹¹. Details regarding the crystallization and refinement of the DIG5.1 (DIG18) structure are provided in the Supplemental Information.

Analysis: Receiver Operator Characteristic (ROC) curves were created for the ranking provided by each group. For each curve, each point indicates the fraction of binding sequences (true positives) versus the fraction of the non-binding sequences (false positives) ranked above each possible ranking threshold. ROC plots are evaluated by the area under the resulting curve (AUC). An AUC of 1 indicates all binding sequences were ranked higher than non-binding sequences, 0.5 indicates the binding sequences were evenly placed in the ranking and 0 indicates all the non-binding sequences were ranked above the binding sequences.

The 95% confidence interval (CI) of the AUC was computed by bootstrapping. For this, a sample set of sixteen sequences and scores was chosen with replacement from the sixteen submitted sequences by a group. The ROC curve was then computed for this sample and the AUC calculated. This was repeated 1000 times. The upper bound to the 95% CI is the 97.5% quantile and the lower bound to the 95% CI is the 2.5% quantile of the resulting distribution of AUCs. The curves and confidence intervals were calculated using code kindly provided by Ajay Jain from UCSF^{21,22}.

The submitted three-dimensional structures for DIG5.1 (DIG18) and DIG10.3 (DIG19) were compared to the experimental crystal structures using the standard RMSD of the backbone C α .²³ The RMSD of all atoms of the active-site residues was also computed between each submitted structure and the respective experimental crystal structure using the ‘rms’ command in PyMol²⁴. The active-site residues are defined to be any residue within 6 Å of the bound digoxigenin molecule in the respective experimental crystal structure.

Phase 2:

Rules: Participants were provided a set of 200 docking decoys each to score for two known structures, DIG5.1(DIG18, PDBID: 5BVB) and DIG10.2(DIG20, PDBID: 4J8T)¹¹. The decoy set included one near-native pose that was within 0.3Å RMSD of the known crystal pose but no other pose was within 2 Å of the known pose. The task was to only rank/score the poses to determine the near-native pose out of the 200. Participants returned their ranked lists or scores for all poses.

Dataset Generation: The ligand decoy set was prepared by performing native docking of the ligand using DOCK version 6.5²⁵. A diverse set of 200 poses for each compound was then selected using the ranking obtained from the “Diverse Subset” utility in MOE 2011.10²⁶ based on conformations, being careful to only include one near native pose for each set. No requirement is made for the RMSD between decoys, only that they provide a diverse subset using the utility in MOE. This greatly decreases the likelihood that the near native pose is the obvious answer based on only pairwise decoy RMSD. Each set of ligands was provided to the participants as a single ‘multi- .mol2’ file with each pose having a unique name.

The structures for DIG18 (DIG5.1) and DIG20 (DIG10.2) were provided as ‘.mol2’ files. The protein structure was set up for scoring using the following protocol. All histidines, asparagines, and glutamines within the active site were examined for the appropriate tautomer at pH 7. The small molecule, crystallographic waters and additives were removed. Residues with multiple orientations within 6 Å of the native ligand were examined, and the appropriate orientation was retained. The ‘A’ conformation, as designated by the alternate location field of the ‘.pdb’ file, was retained for residues outside of 6 Å of the ligand. The N- and C-terminal residues were capped using an acetyl group (ACE) and an N-methyl group (NME), respectively. Hydrogens were added using the default procedure in MOE2011.10. Missing side chains were added for 17 residues in DIG5.1 (DIG18) and 33 residues in DIG10.2 (DIG20). Partial charges were added with the MMFF94x forcefield. The added caps, hydrogens and sidechains were minimized using the MMFF94x forcefield with the default parameters in MOE2011.10 keeping all other atoms fixed. All setup procedures were performed using MOE2011.10²⁶.

Analysis: We consider a successful ranking as one that ranks the correct pose in the top 3. We also analyze the results based solely on the top pose.

Phase 3:

Rules: Participants were given the structure of DIG10.2 and ten ligands to dock. The participants were then asked to rank/score the resulting poses. The binding affinities of these ten small molecules for DIG10.2 were determined experimentally by CSAR, using the Thermofluor²⁷ technique (Methods and details of the Thermofluor assay are given in the Supplemental Information). Affinities were not provided with the structures to ensure a blind study. Although all molecules bound the protein, participants were not told if each small molecule was active. Participants were asked to return the top-3 poses and scores/ranks for each ligand. The structures were given prepared for docking, but participants were free to utilize/setup the structures as necessary for their particular protocol.

Dataset Generation: The protein structure was set up for docking using the same protocol as described in phase 2. The structure was provided as a ‘.mol2’ file. The compounds chosen for the binding affinity experiments were selected by first performing a substructure search of the online Sigma-Aldrich catalog using the four fused-ring, steroid core of digoxigenin. The list was then sorted by price and availability as time and funds were limited. From the pared down list, the ten compounds used in the exercise were chosen to

span chemical functionality at the 3 and 17 positions of the steroid skeleton (Table 2). The molecules chosen are extremely rigid with few rotatable bonds. This limits the amount of conformational space available for the ligand and may remove a potential challenge of docking and scoring. Removing the challenge of a flexible ligand provides the advantage of focusing on the performance of docking and scoring functions on an unknown, designed protein. The small molecules were provided in a '.mol2' format with 3-dimensional coordinates generated using OMEGA^{28,29} to minimize and generate a single, low-energy conformation from the SMILES string. For CS331 and CS332, the hydrogen was removed from the carboxylate group. Participants were allowed to add charges and manipulate the ligand as necessary for their protocol. Binding affinity values used in this analysis were determined using a thermal shift assay (Thermofluor)²⁷. Where possible, binding affinities were also measured using Isothermal titration calorimetry (ITC) (Table 2). The ITC affinities matched those from the thermofluor, but were not distributed to the participants, since only three compounds, CS334, CS335, and CS337, could be determined by this method. The other seven molecules were incompatible with ITC because of weak affinity and/or poor solubility. Details of the ITC experimental methods are provided in the Supplemental Information.

Analysis: To assess how well each method performed on predicting the affinity of the small molecules to DIG10.2, we calculated the Pearson correlation coefficient (r). To determine how well the methods ranked the compounds, the non-parametric correlation coefficients Spearman- ρ and Kendall- τ were calculated. The 95th percent confidence intervals were computed for the Pearson and Spearman correlations using the Fisher transform.³⁰ The 95th percent confidence interval for the Kendall coefficient is approximated by $\tau + 95\% = \tau + \sigma * 1.96$ and $\tau - 95\% = \tau - \sigma * 1.96$, where $\sigma = \left(1 - \tau^2 \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$.³¹ Linear regression was performed on the predicted score versus the experimental affinities, and R^2 values were reported. All correlations and linear regressions were performed using JMP³².

Programs used by participants:

Participating groups provided details on the programs used in generating their results with some groups utilizing modified or custom versions of the programs or scoring functions. Some groups also utilized visual inspection of models or poses to assist in ranking/scoring. Below is a list of the programs used, avoiding using the names of each group to preserve anonymity.

For homology modeling in Phase 1, MUSTER³³ was used by two groups (Group B and Group L); Medusa³⁴ (Group G), Prime^{35,36} (Group D), M4T³⁷ (Group F), Modeller³⁸ (Group J), and GalaxyTBM³⁹ (Group K) were each used by individual groups. Two groups used I-TASSER⁴⁰ (Group E and Group M), which utilizes a consensus of several different threading methods and chooses the best structure. Three groups (Group A, H, and I) declined to provide information about their methods.

For the docking portion of Phase 1 and Phase 3, a wider range of programs were utilized: Glide⁴¹ (Groups A, S and T), Autodock Vina⁴² (Groups B, F, and M-3), MedusaDock⁴³

(Group C), Fred⁴⁴ (Group D), SMINA⁴⁵ (Group E-2-3,14-15), GalaxyDock⁴⁶ (Group K), MDock⁴⁷ (Group J, M, M-3), Surflex-Dock^{21,48} (Group N), and Wilma⁴⁹ (Group O, V).

Scoring was performed in all three phases of the experiments and these scoring functions were used: MM/GBSA⁵⁰ (Group A, G), Autodock Vina⁴² (Group B, E-3,6,9,11-14, F, L, M-3), MedusaScore⁵¹ (Group C), Szybki⁵² (Group D), ConvexPL (Group I, unpublished), ITScore⁵³ (Group J, and M-1-2), PatchSurfer⁵⁴ (Group L), GalaxyDock⁴⁶ (Group K), ICM⁵⁵ (Group N), SIE⁴⁹/SIE-FISH⁵⁶ (Group O), Glide-SP⁴¹ (Group S-1, S-2), Glide-XP⁵⁷ (Group S3, S-4), and a QM/MM linear-response method (Group V).

Results and Discussion:

No two groups submitted the same predictions, despite the use of the same forcefields, programs, or servers. This notable result shows how minor details can significantly alter outcomes. It also underscores the importance of blinded exercises and the subtle bias that is inherent to retrospective studies. When starting a project, there are many small details and parameters to choose. Surely, each participant thought they chose the best options. In a retrospective, exercise, it is too easy to justify changing these parameters if difficulties arise (e.g. “method A is known to have difficulties with this system”, “either option X or Y is a valid choice”, etc.). This can conceal negative outcomes that could be needed to advance our field. Unfortunately, participants did not provide exact details of their calculations with their methods with their submissions, so we cannot identify the causes for different submissions based on the same methods.

Phase 1:

The goal of phase 1 was to determine if computational methods could predict which designed protein sequences were able to bind digoxigenin. Participants had 16 sequences to rank, of which 4 sequences could bind digoxigenin. A total of 12 groups participated in phase 1 of the 2013 CSAR Exercise. One of the groups submitted four different methods, for a grand total of 16 methods. In general, groups utilized sequence-based homology modeling methods to choose a template from the PDB and generate a potential 3-dimensional fold of the protein. The groups would then identify potential pockets to dock and score with their method to determine which sequences bound digoxigenin. All but three of the methods had scores for all of the sequences, one submitted scores for 15 of the 16 sequences, another for 12 of the sequences, and the other for 11 of the sequences. The summary of the performance for each group is displayed in Figure 2.

Almost all groups had active sequences ranked near the top, with a couple of exceptions. The average AUC of the ROC curves for all the submissions was 0.76 with a standard deviation of 0.25. Of the groups that ranked all of the structures, the average AUC is 0.83 with a standard deviation of 0.15. This shows that most groups were able to rank the active sequences before ranking the inactive sequences. All but two of the methods ranked at least one active sequence in the top 3. If only the methods which submitted scores for all sequences are considered, the average rank of the highest scored active sequence is 1.77 (median = 1), and the average rank of the lowest scored active sequence was 7.85 (median = 7). Two methods (Group E-4 and Group H) showed perfect performance in that they were

able to rank the four active sequences as the top four. The rankings and AUC are shown in Figure 2.

The 12 inactive sequences chosen for this phase were designed to make appropriate contacts with digoxigenin¹¹; however, the overall similarity to the active sequences is very low (Table 1). It is possible that similarity between the active sequences was enough to select the appropriate sequences. The four binding sequences had a high degree of similarity to each other and were distinct from the other 12 sequences because the binding sequences all derive from the same scaffold protein. This may have contributed to success in this phase for participants with methods that evaluated sequence similarity. The minimum similarity between pairs of active sequences was 92%; no other pair of sequences had similarity greater than 40%, with the exception DIG6 and DIG17, which were 93% similar to each other. In fact, the best performing strategy (Group E-4) utilized multiple sequence alignments of the dataset followed by clustering of these alignments using a neighbor joining tree to rank the sequences.

Not every group was able to successfully rank the active sequences in the top-three. Groups I and M had poor AUCs (< 0.5). In addition to ranking the protein sequences, 10 of the 13 groups (Groups B-F, H-J, L, and M) submitted the three-dimensional model of the protein used to rank. Group I ranked the active sequences last (AUC = 0), which appears to be due to using a poor model. The RMSD of the active-site residues for their submitted models of DIG5.1 (DIG18) and DIG10.2 (DIG20) were greater than 12 Å, whereas all other groups had active-site RMSDs less than 3 Å. Unfortunately, the group did not provide details on the homology modeling. In investigating their structure, the group did find the correct protein fold; however, the residues in the secondary structure were shifted six residues due to improper modeling of the N-terminus, which forced the binding site to be incorrect. The poor structure makes it difficult to assess the scoring function, since it should fail at determining binding if the protein model is wrong. It is unclear why Group M did not obtain a higher AUC. Group M utilized I-TASSER as did Group E-1–3 for homology modeling and obtained appropriate structures. The issue does not appear to be the scoring function either as they utilized ITScore, which was also used by Group J to score the docked digoxigenin.

Most groups were able to demonstrate success in ranking sequences (AUC $\gg 0.5$), and the majority of computational methods are ranking binding sequences higher than non-binding sequences, at least for the designed protein/digoxigenin system. The confidence intervals around the AUC are large due to a small number of sequences, so comparisons between groups are not statistically significant. The large confidence intervals also make it difficult to assess how well these methods would perform on other protein/ligand complexes. More exercises should be performed in the future as a wide range of designed systems become available through research in the computational protein-design field. Decoy sequences must also have higher similarity to the active sequences.

Phase 2:

The goal of Phase 2 was to determine if a near-native pose could be selected out of a set of decoys. With 199 decoys and 1 near-native pose, it is unlikely that the near-native pose would be randomly selected at the top of a ranked list ($1/200$ or 0.05 %). No actual docking is

performed, so this tests the computational methods ability to score poses. The design of this phase is to decouple the docking problem from the scoring problem, where several poses are generated by an algorithm and the best pose needs to be selected by ranking using a scoring function. The setup of the protein and poses were explained to participants. We emphasized to participants that the poses may not be local minima for all force fields, so short minimizations could be necessary for optimal performance of their method.

A total of 14 groups participated in the second phase of the exercise, with a total of 20 methods. Nine (B-E, H-J, L and M) of the groups had participated in the Phase 1. A summary of where each group docked the near-native poses is found in Table 3. In general, groups/methods were able to rank the near-native pose for each structure as the top-ranking pose or within the top-3 of their rankings. Table 3 provides details of the rankings of the near-native pose. For DIG5.1 (DIG18), 18 of 25 methods were able to rank the near-native pose as the top pose, and 22 of 25 methods were able to rank it in the top-3. DIG10.2 proved to be a little more challenging as only 17 of the 25 methods were able to rank the near-native pose as the top pose, and the other 8 methods did not rank it in the top-3. There is no consensus as to the reason each of these programs did not rank the near-native pose in the top 3. Group P did not provide any computational details while Group D utilized the energy-based scoring approach of Szybki⁵². For the two groups which only succeeded on one of the proteins, Group N used a consensus scoring of the energy-based function from ICM⁵⁸ and the shape-based scoring function from Surflex-Dock²¹. Group R used an in-house generated support vector machine (SVM) with parameters obtained from either Autodock (R-1), Glide (R-4), Liaison (R-5), MMGBSA (R-6) or a combination of Autodock and Glide (R-3) or all four (R-2).

One reason this may have been a particularly easy task is that the protein had only one obvious pocket for the molecule to bind. If there had been other possible clefts in the protein, this would have been a more difficult selection process, testing the scoring functions more rigorously. In these two cases the correct pose is buried in the only possible pocket, whereas all other poses are significantly solvent exposed. This can be seen in Figure 3 which shows the distribution of the decoys in the binding site for both DIG5.1 (DIG18) and DIG10.2 (DIG20). Many force fields have favorable terms for burying hydrophobic atoms in the protein; hence, decoy poses with larger exposed surface area and fewer atomic contacts generally score poorly.

Additionally, a major requirement for exhibiting success with docking protocols such as Autodock⁵⁹, DOCK⁶⁰, FlexX⁶¹, Glide⁴¹, Gold⁶², as well as a host of others, was to reproduce the pose of the small molecule in an X-ray crystal structure to within 2 Å. Many evaluations of these programs on known systems have demonstrated their success on some, but not all, targets^{17,63–67}. These two protein/ligand complexes appear amenable to success in docking programs as only three of the fourteen methods failed to select the near-native pose. With only two proteins and one ligand, it is difficult to discern if the methods would function on other designed protein/ligand complexes.

Phase 3:

CSAR has utilized Thermofluor to obtain equilibrium dissociation constants for ten steroid ligands which were chosen for their potential ability to bind the designed DIG proteins (Table 2). Due to poor solubility of the steroid compounds, we were only able to determine the dissociation constants for three of the ligands using ITC. The two methods agree well for the three compounds with a root-mean square error of 0.23 log units. Phase 3 was designed to evaluate the ability to dock and rank/score a series of 10 steroid ligands towards a designed protein. Fourteen groups participated in phase 3 of the exercise, with six groups submitting multiple methods. The total number of methods submitted was 27. In this phase, the scoring was more difficult, but not impossible; 8 groups (12 total methods) were able to obtain a Pearson correlation coefficient (r) to the experimental affinity of greater than 0.7 with a maximum of 0.9. Of these methods, only 5 of the groups (7 methods) were able to rank with a Spearman correlation (ρ) to the experimental affinity of 0.7. A correlation of 0.7 is chosen as it corresponds to a R^2 of 0.5. For a sample size of only 10 compounds, the 95% confidence interval around these correlation coefficients is very large. The 95% confidence interval around a correlation coefficient of 0.7 with only 10 compounds is 0.13 – 0.92 for Pearson and 0.04 – 0.93 for the Spearman correlation. In previous exercises, we have used molecular weight and SlogP as “null” cases where their correlation to affinity is considered. However, in this exercise, the ligands are all steroids with significant chemical similarity and too small a range in these properties. In general, ranking (Spearman) and correlating (Pearson) scores with binding affinity is a difficult task as has been shown in the previous evaluations of docking and scoring exercises^{17,67,68}. This is also a difficult task due to the narrow range of affinities investigated (pK_d 4.1 to 6.66), with some ligand having affinities within error of each other. All correlations can be seen in Table 4. Comparisons cannot be made between groups because the confidence intervals around the correlation coefficients are large, due to the small number of compounds to rank.

Conclusion:

The 2013 CSAR Benchmark Exercise has been completed. We had a total of 21 different participants using 50 methods across the three phases of the exercise. The results can be best rationalized by considering the energy gaps between correct and incorrect answers in each phase: In the first phase of the exercise, participants successfully determined the four sequences which were able to bind the small molecule digoxigenin, out of a pool of sixteen sequences. The sequences of designed proteins that bind digoxigenin are radically different from those of the proteins that do not bind the small molecule, so it is reasonable to conclude that the energy gaps between the modeled bound poses would be large, and the precision required for discrimination is likely to be low. Therefore, most groups succeeded in this challenge. In the second phase, participants were asked to identify the correct pose from a set of 200 decoy poses. Most decoy poses were considerably more exposed than the correct near-native pose, which suggests the number of 2-body contacts were notably different and the energy gap between the near-native and decoy poses may have been relatively high. Similar to Phase 1, it would be likely that low-precision methods could discriminate between the native and decoy poses. Of course, high-precision methods would do well.

In contrast, the third phase challenge – docking and rank-ordering structurally similar congeneric steroids by affinity – involved much smaller energy gaps between various bound poses of structurally similar compounds. Finer precision of the scoring functions is critical for success. The results for the third phase demonstrate this difficulty, despite the ability of the scoring functions to determine appropriate structures in the first two phases. Nevertheless, qualitative trends are encouraging, and high correlation with experimentally determined affinity could be obtained by a few groups and methods (*e.g.*, E-12). Many participants in this exercise have contributed papers to this special section of the *Journal of Chemical Information and Modeling*. Their detailed analysis of their methods should provide insight into what choices corresponded best with success. The results of the exercise provide a snapshot of the overall state of our ability as a community to model protein-ligand interfaces, and highlight pressing areas where methodological improvements are needed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The authors would like to thank the Chemical Computing Group and OpenEye for their generosity in providing MOE and Omega, respectively. Funding for CSAR and this work is from a U01 grant from NIGMS/NIH (GM086873). CET, SDK, and DB would like to acknowledge funding from DTRA. SDK acknowledges a grant from the NSF (MCB-1330760). We would like to acknowledge Barry Stoddard at the Fred Hutchinson Cancer Research Center for work on the crystallography of DIG5.1 (DIG18).

References

- (1). Yuriev E; Holien J; Ramsland PA Improvements, Trends, and New Ideas in Molecular Docking: 2012–2013 in Review. *J. Mol. Recognit* 2015
- (2). Leach AR; Shoichet BK; Peishoff CE Prediction of Protein–Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem* 2006, 49, 5851–5855. [PubMed: 17004700]
- (3). Sousa SF; Ribeiro AJM; Coimbra JTS; Neves RPP; Martins SA; Moorthy NSHN; Fernandes PA; Ramos MJ Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr. Med. Chem* 2013, 20, 2296–2314. [PubMed: 23531220]
- (4). Babine RE; Bender SL Molecular Recognition of Protein–Ligand Complexes: Applications to Drug Design. *Chem. Rev* 1997, 97, 1359–1472. [PubMed: 11851455]
- (5). Dunbar JB; Smith RD; Yang C-Y; Ung PM-U; Lexa KW; Khazanov NA; Stuckey JA; Wang S; Carlson HA CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *J. Chem. Inf. Model* 2011, 51, 2036–2046. [PubMed: 21728306]
- (6). www.csardock.org.
- (7). Lippow SM; Tidor B Progress in Computational Protein Design. *Curr. Opin. Biotechnol* 2007, 18, 305–311. [PubMed: 17644370]
- (8). Samish I; MacDermaid CM; Perez-Aguilar JM; Saven JG Theoretical and Computational Protein Design. *Annu. Rev. Phys. Chem* 2011, 62, 129–149. [PubMed: 21128762]
- (9). Schreier B; Stumpp C; Wiesner S; Höcker B Computational Design of Ligand Binding Is Not a Solved Problem. *Proc. Natl. Acad. Sci* 2009, 106, 18491–18496. [PubMed: 19833875]
- (10). Leaver-Fay A; Tyka M; Lewis SM; Lange OF; Thompson J; Jacak R; Kaufman K; Renfrew PD; Smith CA; Sheffler W; Davis IW; Cooper S; Treuille A; Mandell DJ; Richter F; Ban Y-EA; Fleishman SJ; Corn JE; Kim DE; Lyskov S; Berrondo M; Mentzer S; Popovi Z; Havranek JJ; Karanickolas J; Das R; Meiler J; Kortemme T; Gray JJ; Kuhlman B; Baker D; Bradley P Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol* 2011, 487, 545–574. [PubMed: 21187238]

- (11). Tinberg CE; Khare SD; Dou J; Doyle L; Nelson JW; Schena A; Jankowski W; Kalodimos CG; Johnsson K; Stoddard BL; Baker D Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. *Nature* 2013, 501, 212–216. [PubMed: 24005320]
- (12). Moulton J; Fidelis K; Kryzhanovskiy A; Schwede T; Tramontano A Critical Assessment of Methods of Protein Structure Prediction (CASP) — Round X. *Proteins Struct. Funct. Bioinforma* 2014, 82, 1–6.
- (13). Lensink MF; Wodak SJ Docking and Scoring Protein Interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinforma* 2010, 78, 3073–3084.
- (14). Guthrie JP SAMPL4, a Blind Challenge for Computational Solvation Free Energies: The Compounds Considered. *J. Comput. Aided Mol. Des* 2014, 28, 151–168. [PubMed: 24706106]
- (15). Dunbar JB Jr; Smith RD; Damm-Ganamet KL; Ahmed A; Esposito EX; Delproposto J; Chinnaswamy K; Kang Y-N; Kubish G; Gestwicki JE; others CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model* 2013, 53, 1842–1852. [PubMed: 23617227]
- (16). Dunbar JB Jr; Smith RD; Yang C-Y; Ung PM-U; Lexa KW; Khazanov NA; Stuckey JA; Wang S; Carlson HA CSAR Benchmark Exercise of 2010: Selection of the Protein–ligand Complexes. *J. Chem. Inf. Model* 2011, 51, 2036–2046. [PubMed: 21728306]
- (17). Warren GL; Andrews CW; Capelli A-M; Clarke B; LaLonde J; Lambert MH; Lindvall M; Nevins N; Semus SF; Senger S; Tedesco G; Wall ID; Woolven JM; Peishoff CE; Head MS A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem* 2006, 49, 5912–5931. [PubMed: 17004707]
- (18). Jansen JM; Cornell W; Tseng YJ; Amaro RE Teach–Discover–Treat (TDT): Collaborative Computational Drug Discovery for Neglected Diseases. *J. Mol. Graph. Model* 2012, 38, 360–362. [PubMed: 23085175]
- (19). <https://www.kaggle.com/c/MerckActivity>.
- (20). Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE The Protein Data Bank. *Nucleic Acids Res* 2000, 28, 235–242. [PubMed: 10592235]
- (21). Jain AN Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput. Aided Mol. Des* 2007, 21, 281–306. [PubMed: 17387436]
- (22). Lei S; Smith MR Evaluation of Several Nonparametric Bootstrap Methods to Estimate Confidence Intervals for Software Metrics. *IEEE Trans. Softw. Eng* 2003, 29, 996–1004.
- (23). Damm KL; Carlson HA Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures. *Biophys. J* 2006, 90, 4558–4573. [PubMed: 16565070]
- (24). Schrodinger LLC. The PyMol Molecular Graphics System
- (25). Lang PT; Brozell SR; Mukherjee S; Pettersen EF; Meng EC; Thomas V; Rizzo RC; Case DA; James TL; Kuntz ID DOCK 6: Combining Techniques to Model RNA–small Molecule Complexes. *RNA* 2009, 15, 1219–1230. [PubMed: 19369428]
- (26). Chemical Computing Group Inc. MOE2011.10; 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011.
- (27). Pantoliano MW; Petrella EC; Kwasnoski JD; Lobanov VS; Myslik J; Graf E; Carver T; Asel E; Springer BA; Lane P; Salemme FR High-Density Miniaturized Thermal Shift Assays as a General Strategy for Drug Discovery. *J. Biomol. Screen* 2001, 6, 429–440. [PubMed: 11788061]
- (28). Hawkins PCD; Skillman AG; Warren GL; Ellingson BA; Stahl MT OMEGA 2.5.1.4; OpenEye Scientific Software, Santa Fe, NM.
- (29). Hawkins PCD; Skillman AG; Warren GL; Ellingson BA; Stahl MT Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model* 2010, 50, 572–584. [PubMed: 20235588]
- (30). Fisher RA On the “Probable Error” of a Coefficient of Correlation Deduced from a Small Sample 1921.
- (31). Bonett DG; Wright TA Sample Size Requirements for Estimating Pearson, Kendall and Spearman Correlations. *Psychometrika* 2000, 65, 23–28.

- (32). JMP Pro 10; SAS Institute, Inc.: Cary, N.C.
- (33). Wu S; Zhang Y MUSTER: Improving Protein Sequence Profile–profile Alignments by Using Multiple Sources of Structure Information. *Proteins Struct. Funct. Bioinforma* 2008, 72, 547–556.
- (34). Ding F; Dokholyan NV Emergence of Protein Fold Families through Rational Design. *PLoS Comput Biol* 2006, 2, e85. [PubMed: 16839198]
- (35). Jacobson MP; Pincus DL; Rapp CS; Day TJF; Honig B; Shaw DE; Friesner RA A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Bioinforma* 2004, 55, 351–367.
- (36). Jacobson MP; Friesner RA; Xiang Z; Honig B On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol* 2002, 320, 597–608. [PubMed: 12096912]
- (37). Fernandez-Fuentes N; Madrid-Aliste CJ; Rai BK; Fajardo JE; Fiser A M4T: A Comparative Protein Structure Modeling Server. *Nucleic Acids Res* 2007, 35 (suppl 2), W363–W368. [PubMed: 17517764]
- (38). Fiser A; Šali A Modeller: Generation and Refinement of Homology-Based Protein Structure Models. In *Methods in Enzymology*; Charles W. Carter J and R. M. S, Ed.; Macromolecular Crystallography, Part D; Academic Press, 2003; Vol. 374, pp 461–491. [PubMed: 14696385]
- (39). Ko J; Park H; Seok C GalaxyTBM: Template-Based Modeling by Building a Reliable Core and Refining Unreliable Local Regions. *BMC Bioinformatics* 2012, 13, 198. [PubMed: 22883815]
- (40). Roy A; Kucukural A; Zhang Y I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc* 2010, 5, 725–738. [PubMed: 20360767]
- (41). Friesner RA; Banks JL; Murphy RB; Halgren TA; Klicic JJ; Mainz DT; Repasky MP; Knoll EH; Shelley M; Perry JK; Shaw DE; Francis P; Shenkin PS Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem* 2004, 47, 1739–1749. [PubMed: 15027865]
- (42). 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]
- (43). Ding F; Yin S; Dokholyan NV Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *J. Chem. Inf. Model* 2010, 50, 1623–1632. [PubMed: 20712341]
- (44). McGann MR; Almond HR; Nicholls A; Grant JA; Brown FK Gaussian Docking Functions. *Biopolymers* 2003, 68, 76–90. [PubMed: 12579581]
- (45). Koes DR; Baumgartner MP; Camacho CJ Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model* 2013, 53, 1893–1904. [PubMed: 23379370]
- (46). Shin W-H; Seok C GalaxyDock: Protein–Ligand Docking with Flexible Protein Side-Chains. *J. Chem. Inf. Model* 2012, 52, 3225–3232. [PubMed: 23198780]
- (47). Huang S-Y; Zou X Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins Struct. Funct. Bioinforma* 2007, 66, 399–421.
- (48). Spitzer R; Jain AN Surflex-Dock: Docking Benchmarks and Real-World Application. *J. Comput. Aided Mol. Des* 2012, 26, 687–699. [PubMed: 22569590]
- (49). Sulea T; Hogues H; Purisima EO Exhaustive Search and Solvated Interaction Energy (SIE) for Virtual Screening and Affinity Prediction. *J. Comput. Aided Mol. Des* 2011, 26, 617–633. [PubMed: 22198519]
- (50). Massova I; Kollman PA Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des* 2000, 18, 113–135.
- (51). Yin S; Biedermannova L; Vondrasek J; Dokholyan NV MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model* 2008, 48, 1656–1662. [PubMed: 18672869]
- (52). SZYBKI 1.8.0.2; OpenEye Scientific Software, Santa Fe, NM.

- (53). Huang S-Y; Zou X An Iterative Knowledge-Based Scoring Function to Predict Protein–ligand Interactions: II. Validation of the Scoring Function. *J. Comput. Chem* 2006, 27, 1876–1882. [PubMed: 16983671]
- (54). Hu B; Zhu X; Monroe L; Bures MG; Kihara D PL-PatchSurfer: A Novel Molecular Local Surface-Based Method for Exploring Protein-Ligand Interactions. *Int. J. Mol. Sci* 2014, 15, 15122–15145. [PubMed: 25167137]
- (55). Abagyan R; Totrov M; Kuznetsov D ICM—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem* 1994, 15, 488–506.
- (56). Corbeil CR; Sulea T; Purisima EO Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model. *J. Chem. Theory Comput* 2010, 6, 1622–1637. [PubMed: 26615695]
- (57). Friesner RA; Murphy RB; Repasky MP; Frye LL; Greenwood JR; Halgren TA; Sanschagrin PC; Mainz DT Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *J. Med. Chem* 2006, 49, 6177–6196. [PubMed: 17034125]
- (58). Abagyan R; Totrov M; Kuznetsov D ICM—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem* 1994, 15, 488–506.
- (59). Park H; Lee J; Lee S Critical Assessment of the Automated AutoDock as a New Docking Tool for Virtual Screening. *Proteins Struct. Funct. Bioinforma* 2006, 65, 549–554.
- (60). Brozell SR; Mukherjee S; Balias TE; Roe DR; Case DA; Rizzo RC Evaluation of DOCK 6 as a Pose Generation and Database Enrichment Tool. *J. Comput. Aided Mol. Des* 2012, 26, 749–773. [PubMed: 22569593]
- (61). Kramer B; Rarey M; Lengauer T Evaluation of the FLEXX Incremental Construction Algorithm for Protein–ligand Docking. *Proteins Struct. Funct. Bioinforma* 1999, 37, 228–241.
- (62). Verdonk ML; Cole JC; Hartshorn MJ; Murray CW; Taylor RD Improved Protein–ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma* 2003, 52, 609–623.
- (63). Wang R; Lu Y; Wang S Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem* 2003, 46, 2287–2303. [PubMed: 12773034]
- (64). Kellenberger E; Rodrigo J; Muller P; Rognan D Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins Struct. Funct. Bioinforma* 2004, 57, 225–242.
- (65). Kontoyianni M; McClellan LM; Sokol GS Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem* 2004, 47, 558–565. [PubMed: 14736237]
- (66). Li X; Li Y; Cheng T; Liu Z; Wang R Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J. Comput. Chem* 2010, 31, 2109–2125. [PubMed: 20127741]
- (67). Damm-Ganamet KL; Smith RD; Dunbar JB; Stuckey JA; Carlson HA CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model* 2013, 53, 1853–1870. [PubMed: 23548044]
- (68). Smith RD; Dunbar JB; Ung PM-U; Esposito EX; Yang C-Y; Wang S; Carlson HA CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *J. Chem. Inf. Model* 2011, 51, 2115–2131. [PubMed: 21809884]

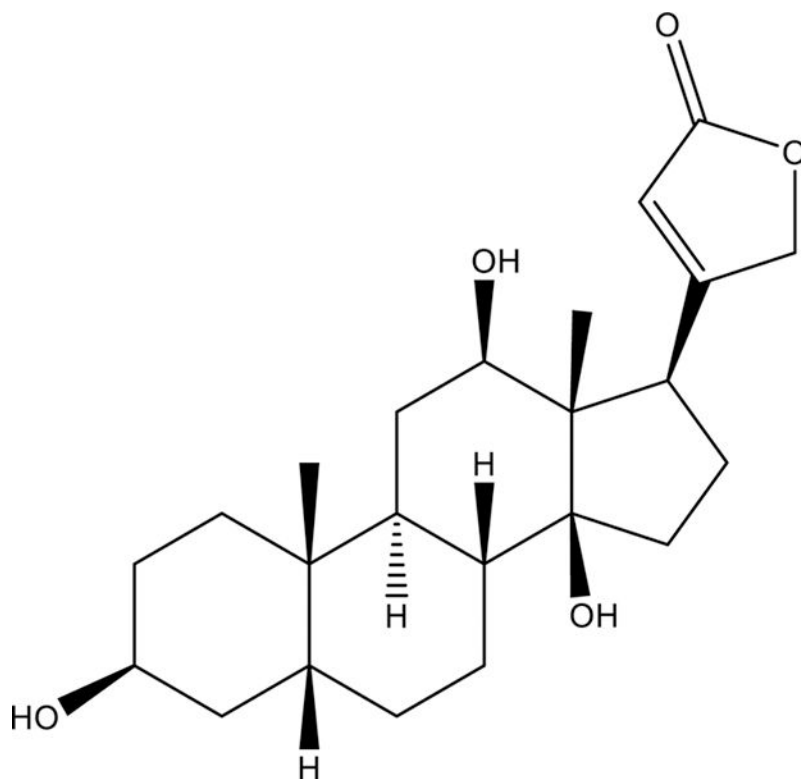


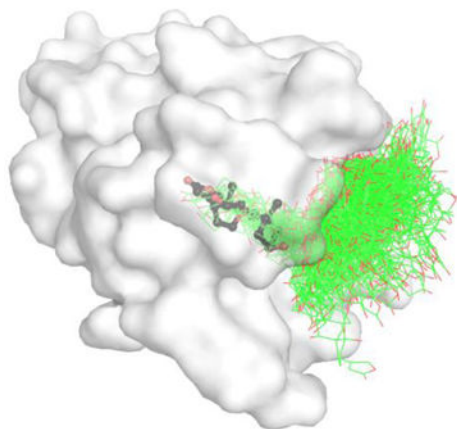
Figure 1.
Structure of digoxigenin (CS337).

DIGID	Group															
	E-4	H	E-1	C	B	L	E-3	F	K	J	E-2	D	G	A	M	I
DIG1	15	5	6	16	16	16	2	10	8	2	16	14	15		4	1
DIG2	10	5	16	8	7	4	16	1	8	15	14	14	7	5	2	7
DIG3	16	5	15	4	8	5	3	2	8	14	3	11	10	1	3	5
DIG4	5	5	8	7	12	7	11	9	3	12	15	7	11	6	6	4
DIG5	1	3	1	6	1	1	4	4	4	4	1	12	3	2	5	11
DIG6	9	5	4	11	10	9	9	10	8	7	2	13	14	3	7	
DIG7	7	5	7	15	11	14	15	10	8	1	13	16	12		14	8
DIG8	14	5	10	5	2	6	5	6	5	8	5	1	2		1	6
DIG9	12	5	14	14	6	13	12	10	8	11	12	8	13		8	
DIG10	3	1	3	3	4	3	1	5	2	9	6	4	8	4	9	
DIG12	6	5	11	12	13	15	14	10	8	6	8	6	9	11	13	
DIG13	13	5	13	9	9	10	10	8	8	16	9	10		10	16	2
DIG14	11	5	12	13	15	11	13	10	8	13	7	9	4	12	15	3
DIG17	8	5	9	10	14	8	7	10	1	10	10	3	1	9	11	
DIG18	2	2	2	1	5	2	6	3	7	3	4	2	5	7	12	9
DIG19	4	4	5	2	3	12	8	7	6	5	11	5	6	8	10	10
AUC	1.00	1.00	0.98	0.96	0.94	0.83	0.81	0.81	0.81	0.77	0.75	0.73	0.73	0.66	0.46	0.00
95 % CI	1.00 -	1.00 -	0.88 -	0.83 -	0.81 -	0.44 -	0.56 -	0.58 -	0.58 -	0.52 -	0.44 -	0.42 -	0.45 -	0.31 -	0.17 -	0.00 -
1 σ CI	1.00 -	1.00 -	0.96 -	0.92 -	0.88 -	0.67 -	0.71 -	0.69 -	0.71 -	0.67 -	0.6 -	0.58 -	0.61 -	0.50 -	0.33 -	0.00 -
	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.92	0.92	0.90	0.90	0.86	0.81	0.60	0.00

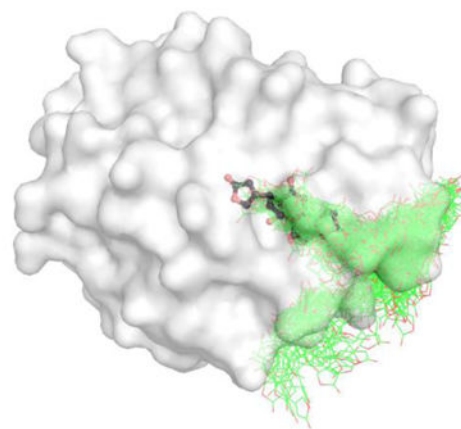
Figure 2.

The rankings are given for each group's predictions of the binding of digoxigenin to 16 sequences. The entries in bold indicate the protein sequences which bound digoxigenin. Boxes in blue were predicted in the top-4, since there are 4 active sequences. Boxes in grey were predicted inactive (outside of top-4). Entries with no color were not ranked by the group. The AUC of the ROC curve is given with the 95% and 1 σ confidence intervals based on 1000 bootstrap samples. The groups are sorted by the AUC. The numbers indicate each group's rank of each sequence.

A. DIG5.1 (DIG18)



B. DIG10.2 (DIG20)

**Figure 3.**

One near-native pose and 199 decoy poses were given for both DIG5.1 (DIG18) (A) and DIG 10.2 (DIG20) (B). The near-native digoxigenin pose is colored with carbon as magenta. The protein surface was computed in PyMol²⁴ and displayed white. The non-native digoxigenin decoys are colored with gray carbons.

Table 1.

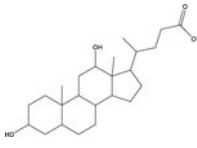
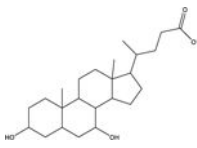
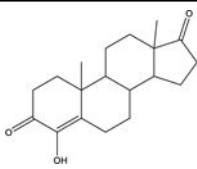
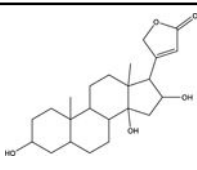
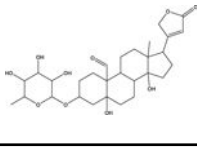
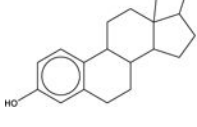
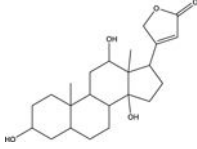
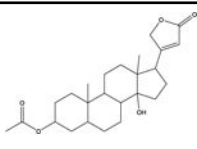
Sequences used for Phase 1 of the exercise. Bolded entries indicate sequences which were able to bind digoxigenin.

ID	Sequence
DIG1	MSLDFNTLAQNFTQFWYNQFDTDRSQLGNLYRNESMTTHETSQLQGAKDIVEGLVSLPFQKYQARI TTLDAQPASPYGDVLVMTGDGLTDEEQNPWRYSAVYHLIPDGNSSYYVFNAIWRYNYSAGS
DIG2	MVSAKDFSGANLYTLEEVQYGKFEARMKMAAASGTVSSMYLHQNGSWIADGRPFVAVYIIVLGKNP GSFQSNIVTGKAGAYKTSAKHHAVSPAADQAFHTYGLEWTPNYVRWTVVDGQEVKTEGGQVSNLT GTQGLRFSLYSSESAAWVGQFDESCLPLFQFINWVKVYKYTPQGEGGSDFTLDWTDNFDTFDGS RWGKGDTYFDGNRVDFTDKNIYSRDGMLILALTRKGQESFNGQVPRDDEPAPQSSSSAPASSGS
DIG3	MGTTPNSTGWHHDGYYYHWWSDGGGDSTYTNNSSGGTYEITWGNNGGILIGGKGWNPGLNARAIHFT GVYQPNGTSLSVYGWTRNPLVSYIIVENFGSSNPSSGSTDLGTVSCDGSYTLGQSTWYNPSID GTQTFNAYWSVRQDKRSSGTVQTGCHFDAWASAGLNVTDGHYYQIVATFGWYSSGYARITVADVG GS
DIG4	MSDVESENTSENRAQVAARQHNRKIVEQYMHTRGEAELKMHLFTEDGVGGSWTSSGQPIAIRG REKLGEHDVFLQVFPDWVWTDIQIFETQDPNFWVECRGEGAIVFPGYPRGQYRAHYLASFRFEN GLIKETRWFVWNPCEAFRALGIEGS
DIG5	MNAKEILVHSLRLENGDARGWCDFHPEGVLEFPYAPPGWKTRFEGRETIWAHMRLHPEHVTVR FTDVQFYETADPDLAIGEYHGDGVVTVSGGKYAADFITVLRTRDQGILLYRVFWNPLRALEAAGGV EAAAKIVQGAGS
DIG6	MNLQTDQTTTTADESAIRAFTRQ MIDAWNRS GEGFAAPFSETADYITFDGTHLKGKKEIAAFAQQA FDTVAKGTRHEGEVDFVRVNSQLALMLTVWRVILPGQTETSASMDALPLYVVTGKDEGWQIEGLLA TYKLT LERGSFLDDFDSLAEARQVTDLVASLKQSHGS
DIG7	MSEPVPFTEAAEDAFYAALAEAGSLDDYMAVWARDHVAFIHPLAAPLNGRAAVAAGWRSHLGAA GRFRLQVKAVHEIRQADHVIRITDIFTGGDETAPRPAALATAVYRREADGWRMVLYHASPLQVGAK AGADTPPVVFHGS
DIG8	MTIAEIAKDYTELNKQGDQAGAYEKYAADDIAYYQAMEGPMVASHGKEAWRQALQWYQENAEFHG GSVEGPYVNGDQFALRFKWDVTPKATGERVTVDGVHLYTVKNGKITEVRWYYGS
DIG9	MKLCFNEFTTLENSNLKLDLELCEKHGYDYIQTMDKLPEYLDHSLDDLAEYFQTHHIKPLALLHLL FFNNRDEKGNHEITEFGKMMETCKTLGVKYVLAWPLRTEQKIVKEEIKSSVDVLTSLDIAEPYGVK IALNFGGLPQCTVNTFEQAYEIVNTVNRDNVGLYLSNFHFHAMGNSIESLKQADGKKIFYGIGDTEDF PIGFLTSEDVMVPGQGAIDLDAHLSALKEIGFSDVVSALLRPEYYKLTAEEAIQTAKKTTVDVVSKEYF SMGS
DIG10	MNAKEIVVHSLRLENGDARGWCDFHPEGVLEYPYAPPGHKTRFEGRETIWAHMRLFPEYVTVR FTDVQFYETADPDLAIGEFHGDGVHTVSGGKLAADYISVLRTRDQGILLYRVFFNPLRVLEALGGVE AAAKIVQGAGS
DIG12	MGMEVNPQDIVAQVQAAFVEYERALVENDIEAMNALFWHTPETVFYGATTVQHGGGEAWRAHVERS QPHPKSRKLHRTVTVTGTDFATVSTFTSDGTPLLRQMQTWARLSPADGWKIVAAHFSLIAMPG S
DIG13	MKLVAGLSSPEELEAEKADVTLHIDLDFSGARVDKEKGLTCMRVSDGGKFEGDERERIEKMKRA FDSLNPDYVYLESDLPSAFDFNCRIAEFYGNVIRTPDYSELKGIVEGRRGDLVVIATMGKSKRDVETI VRILTNYDDVLAHLMG ERFSTFMVLAAYLGSPWICYVVGSPKFPGAISLDDAREIISRLGGS
DIG14	MTLRAARPEFLDLFPAGAEARRLADGFTWTS GPVYVPARSAVIFSDSAQNRTWAWSDDGQLSPEM HPSHHQGGHCLNKQGHLIACSHGLRRLERQREPGGEWESIADSFEGKKNLNSPCLAPDGLWLF SDPTWGIDLPEFGYGGEMELPGRWVFR LAPDGTLSAPIRDRVKPTGLAFLPSGNLLVSDAGDNATH RYCLNARGETEYQG VHFVTEPGATY YLRVDAGGLI WASAGDGVHVLTPDGDDELGRVLTPTTTGLC FGGPEGRTLYMTVSTEFWSIETNVRGGS
DIG17	MNLQTDQTTTTADESAIRAFHRLIDAFNRGSGEGFAAPFSETADFITAEGLHKGKKEIAAYHQAF DTVVKGTRLEGEVDFVRVNSQLALMLVSRILPGQTETSASRDYLPYVVTGKDEGWQIEGLLATR KTLERQFFLDDFDSLAEARQVTDLVASLKQSHGS
DIG5.1 (DIG18)	MNAKEILVHSLRLENGDARGWCDFHPEGVLEYPYAPPGWKTRFEGRETIWAHMRLHPEHVTW RFTDVQFYETADPDLAIGEYHGDGVVTVSGGKYAADYITVLRTRDQGILLRVFWNPLRILEAAGGV EAAAKIVQGAGS
DIG10.3 (DIG19)	MNAKEIVVHALRLENGDARGWSDLFHPEGVLEYPYPPPGYKTRFEGRETIWAHMRLFPEYMTIRF TDVQFYETADPDLAIGEFHGDGVLTASGGKLAYDYIAVWRTRDQGILLYRFFNPLRVLEPLGGVE AAAKIVQGAGS

ID	Sequence
DIG10.2 (DIG20)	MNAKEIVVHALRLLLENGDARGWCDLFHPEGVLEYPYPPPGYKTRFEGRETIWAHMRLFPEYMTIRF TDVQFYETADPDLAIGEFHGDGVHTVSGGKLAADYISVLRTDGGQILLYRLFFNPLRVLEPLGLE (Phase 2 only)

Table 2.

Binding affinities of ligands used in Phase 3.

CSAR ID	Structure	Thermofluor (K_d (error) μM)	ITC (K_d (error) μM)	Thermoluor (pK_d)	ITC (pK_d)
CS331		44.5 (5.9)		4.35	
CS332		46.2 (4.4)		4.34	
CS333		30.6 (7.8)		4.51	
CS334		3.25 (0.34)	2.30 (0.27)	5.49	5.64
CS335		3.12 (0.25)		5.51	
CS336		19.0 (3.4)	7.43 (2.74)	4.72	5.13
CS337 (digoxigenin)		0.474 (0.068)	0.498 (0.027)	6.32	6.30
CS338		0.220 (0.085)		6.66	

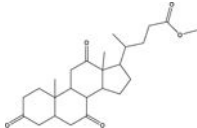
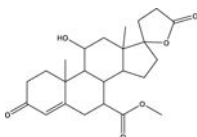
CSAR ID	Structure	Thermofluor (K_d (error) μ M)	ITC (K_d (error) μ M)	Thermoluor (pK_d)	ITC (pK_d)
CS339		79.1 (7.5)		4.10	
CS340		53.7 (3.6)		4.27	

Table 3.

Rankings of the near-native poses from Phase 2.

Group	DIG5.1 (DIG18) rank	DIG10.2 (DIG20) rank
B	1	1
C	1	1
E-5	1	1
E-6	1	1
E-8	1	1
E-9	1	1
E-10	1	1
H	1	1
I	1	1
J	1	1
K	1	1
L	1	1
M	1	1
O-1	1	1
O-2	1	1
E-7	2	1
R-2	2	6
R-5	2	13
R-6	2	13
R-3	1	20
R-1	1	23
N	24	1
R-4	1	27
P	86	12
D	93	36
Ranked near-native best	18 of 25	17 of 25
Ranked near-native Top-3	22 of 25	17 of 25

Table 4.

Correlation coefficients for Phase 3 of the exercise.

Group-Method	R ²	Pearson (r)	r 95% CI	Spearman (ρ)	ρ 95% CI
E-12	0.814	0.902	0.631 – 0.977	0.855	0.386 – 0.973
D	0.766	0.875	0.548 – 0.970	0.709	0.057 – 0.937
S-2	0.751	0.867	0.522 – 0.968	0.746	0.125 – 0.947
T	0.732	0.855	0.489 – 0.965	0.673	–0.005 – 0.927
E-14	0.723	0.850	0.475 – 0.964	0.758	0.149 – 0.950
B	0.671	0.819	0.391 – 0.956	0.554	–0.170 – 0.890
O-1	0.655	0.809	0.366 – 0.953	0.649	–0.042 – 0.920
C-3	0.644	0.802	0.349 – 0.951	0.770	0.174 – 0.953
M-3	0.630	0.794	0.328 – 0.949	0.509	–0.222 – 0.874
O-3	0.609	0.780	0.296 – 0.945	0.576	–0.143 – 0.897
M-2	0.580	0.762	0.254 – 0.940	0.576	–0.143 – 0.897
E-11	0.534	0.731	0.187 – 0.932	0.685	0.015 – 0.930
S-1	0.464	0.681	0.090 – 0.917	0.636	–0.060 – 0.916
M-1	0.459	0.677	0.083 – 0.916	0.406	–0.327 – 0.834
O-4	0.444	0.666	0.063 – 0.913	0.673	–0.005 – 0.927
S-3	0.444	0.666	0.063 – 0.913	0.515	–0.215 – 0.876
U	0.343	0.586	–0.069 – 0.888	0.782	0.201 – 0.956
S-4	0.312	0.559	–0.110 – 0.879	0.479	–0.255 – 0.863
J	0.301	0.548	–0.124 – 0.876	0.612	–0.095 – 0.908
H	0.233	0.483	–0.211 – 0.853	0.273	–0.442 – 0.776
I	0.215	0.464	–0.235 – 0.846	0.600	–0.111 – 0.905
C-2	0.186	0.432	–0.272 – 0.835	0.321	–0.402 – 0.798
L	0.119	0.345	–0.364 – 0.801	0.382	–0.350 – 0.824
V	0.059	0.242	–0.457 – 0.757	0.103	–0.564 – 0.689
E-13	0.043	0.208	–0.486 – 0.74	–0.006	–0.633 – 0.626
C-1	0.017	0.130	–0.544 – 0.702	–0.224	–0.752 – 0.479
E-15	0.004	0.060	–0.592 – 0.664	–0.224	–0.752 – 0.479