JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys

James B. Dunbar, Jr.,*[,†] Richard D. Smith,[†] Kelly L. Damm-Ganamet,[†] Aqeel Ahmed,[†] Emilio Xavier Esposito,[†,‡] James Delproposto,[§] Krishnapriya Chinnaswamy,[§] You-Na Kang,[§] Ginger Kubish,[§] Jason E. Gestwicki,[§] Jeanne A. Stuckey,[§] and Heather A. Carlson*[,†]

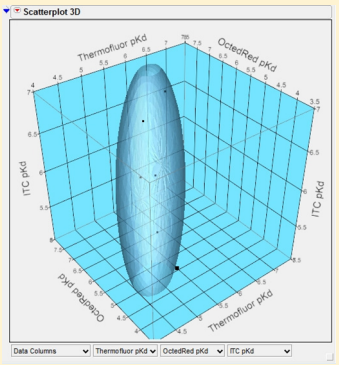
[†]Department of Medicinal Chemistry, University of Michigan, 428 Church St., Ann Arbor, Michigan 48109-1065, United States

[‡] exeResearch LLC, 32 University Drive, East Lansing, Michigan 48823, United States

[§]Center for Structural Biology, University of Michigan, 3358E Life Sciences Institute, 210 Washtenaw Ave., Ann Arbor, Michigan 48109-2216, United States

Supporting Information

ABSTRACT: A major goal in drug design is the improvement of computational methods for docking and scoring. The Community Structure Activity Resource (CSAR) has collected several data sets from industry and added in-house data sets that may be used for this purpose (www.csardock.org). CSAR has currently obtained data from Abbott, GlaxoSmithKline, and Vertex and is working on obtaining data from several others. Combined with our in-house projects, we are providing a data set consisting of 6 protein targets, 647 compounds with biological affinities, and 82 crystal structures. Multiple congeneric series are available for several targets with a few representative crystal structures of each of the series. These series generally contain a few inactive compounds, usually not available in the literature, to provide an upper bound to the affinity range. The affinity ranges are typically 3−4 orders of magnitude per series. For our in-house projects, we have had compounds synthesized for biological testing. Affinities were measured by Thermofluor, Octet RED, and isothermal titration calorimetry for the most soluble. This allows the direct comparison of the biological affinities for those compounds, providing a measure of the variance in the experimental affinity. It appears that there can be considerable variance in the absolute value of the affinity, making the prediction of the absolute value ill-defined. However, the relative rankings within the methods are much better, and this fits with the observation that predicting relative ranking is a more tractable problem computationally. For those in-house compounds, we also have measured the following physical properties: logD, logP, thermodynamic solubility, and pK_a. This data set also provides a substantial decoy set for each target consisting of diverse conformations covering the entire active site for all of the 58 CSAR-quality crystal structures. The CSAR data sets (CSAR-NRC HiQ and the 2012 release) provide substantial, publically available, curated data sets for use in parametrizing and validating docking and scoring methods.



INTRODUCTION

The Community Structure Activity Resource (CSAR) was created to provide better, more reliable, and consistent data that will allow the scientific community to improve their tools for docking and scoring.[1−4] In engineering,[5,6] there are thousands of tables of data, meticulously created, that allow engineers to design on paper (or computer) the vast array of items we see today. Computer-aided design uses this data to correctly design power plants, bridges, packaging material, cars, planes, etc. However, the docking and scoring community cannot provide the same accuracy because we lack the appropriate data.

To help provide for this critical need, CSAR is gathering data from industrial sources augmented with data from the literature and academic laboratories. CSAR has currently obtained data from Abbott, GlaxoSmithKline, and Vertex and is working to obtain data from several others. We are most interested in two types of data sets:

1. Comprehensive sets: Targets with compounds that span 3−8 orders of magnitude in binding (K_i, K_a, K_d, or IC_{50}, no percent inhibition data) also with at least one crystal structure for the series. Some inactives in the series are also needed to help complete the compound set.
2. Activity cliffs (ideally within the comprehensive sets): Pairs of compounds where minimal changes in a ligand result in dramatic changes in binding (either a loss of activity or significant improvement) with affinity data and a crystal structure in the series of the pairs of compounds. These pairs would be characterized by a change of ∼3 or more in the pIC_{50} (or pK_i, etc.) resulting from a change in 1−3 non-hydrogen atoms.

Special Issue: 2012 CSAR Benchmark Exercise

Received: January 21, 2013
Published: April 25, 2013

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

Another goal of CSAR is to help establish appropriate statistical evaluations and incorporate appropriate limits (error bars) for the biophysical measurements. How should one compare the predicted affinity for a given method (or across computational methods) to the experimental data? What are the appropriate statistical methods to employ? How should one compare a docking pose to a crystal structure?

A third goal is to provide a standard for determining what constitutes a quality data set. Affinities cannot be predicted to a precision greater than that of the actual experimental measurements. One of the characteristics of a high quality data set is to include multiple measures of affinity from multiple techniques to provide a means of assessing the overall variance in the affinities. For data sets worked on in-house by CSAR, each individual affinity has the standard error of measurement included. The affinity values are the average of at least three individual measurements and often more. Conditions used across multiple biophysical methods are kept as consistent as possible. This includes items such as pH, buffer, DMSO concentration, protein sequence, and personnel. This removes as many sources of variation as possible. The physical properties of the small molecules were measured by the sole source, WuxiAppTec. These are thermodynamic solubility, logP (pH 11), logD (pH 7.4), and pK_a for the compounds used in the in-house biophysical assays.

As part of its function, the CSAR center also runs benchmarking exercises to aid in stimulating growth in the field. This year, the 2012 Benchmark Exercise is based on four targets: Urokinase, Chk1, ERK2, and LpxC (*Pseudomonas aeruginosa*). The exercise was conducted in two parts: pose prediction of blinded crystal data and predicting rank-order affinity. Each system had a range of affinities, and all but ERK2 contained known inactives.

Previously, we investigated how the current publically available crystal structures could work toward achieving these goals. To this end, we created the CSAR-NRC HiQ data set that is comprised of 343 complexes from the PDB[7,8] as of 2008. This was the origin of the crystal quality metrics currently employed by CSAR and has extensive input from Gregory Warren (Openeye Scientific Software) and Traian Sulea (National Research Council Canada). This data set has been vetted by many organizations worldwide, and any appropriate modifications have been implemented. All of the entries have been setup in a consistent fashion, ready for docking studies. The CSAR-NRC HiQ set contains 52 protein targets with 2 or more structures (9 targets have 4 or more entries) and 191 targets with a single entry. We are currently updating the set to include appropriate PDB structures from 2008 to 2011.

There are a few comparable crystallography-based data sets found in the literature. One is from CCDC/Astex[9] comprised of only 85 diverse protein−ligand PDB complexes that have drug-like small molecules, and all targets are represented only once. This data set is not currently being augmented or updated. A new data set generated by Openeye called Iridium[10] has just been released. It began as the compilation of several literature sets (includes GlideXP,[11] CCDC/Astex,[9] Vertex set,[12] Gold[13]) numbering 728 PDB complexes. This was reduced to a set of 233 complexes by requiring structure factor data be available. Of these, 121 were considered Highly Trustworthy, 104 were considered Mildly Trustworthy, and the remainder (8) are considered Untrustworthy. The criteria for the Highly Trustworthy are very similar to the requirements of the CSAR. SERAPhic[14] is another data set aimed at the niche of fragment-based design work. It is a very small data set of 53 complexes from the existing literature. All of these data sets are, to our knowledge, currently static or being reduced in size.

Another aspect of quality data is the variance in the experimental data itself. Stouch[15] has commented in a recent paper on the error in calculated energies, which could be as large as the entire useable range of drug discovery. He highlights the need to understand this error and provide confidence limits for the output. These concepts need to be extended to include a similar analysis of variation in the experimental affinities—what really is the "true" affinity (value ± error) that we are trying to predict? A recent paper by Shivakumar[16] et al. has found that even using the newest version of OPLS (2.0), 26 of 239 compounds have an average unsigned error (AUE) of between 1.0 and 1.2 kcal/mol in the calculation of the absolute solvation free energies, and around 100 compounds have an AUE between ∼0.6 and 1.2 kcal/mol. There is 0.5 to 1 pK unit of error for a major portion of the set. As this is only one part of the complete free energy (ΔG_{bind}) of binding, the AUE should increase to 1 or greater pK units. Historically scientists have worked with a single affinity value and not taken into account the variance that is associated with the measured value. Recently, there has been more attention paid to estimating and understanding what the variance in the affinity[17−19] could be, and the best predictions that could be done may be to ∼1 pK unit of variance.

In analyzing the ChEMBL[20] database for multiple ligand target affinities, Kramer[19] et.al. have found 2540 complexes with at least two affinity measurements. Of these, 1699 target ligand measurements are within 1 pK unit of one another. This is only ∼0.53% of the 320,520 existing published affinities in ChEMBL at that time. For the 841 remaining complexes, we may never know why the affinity values differ greatly. This may be caused by different assays, different conditions, different sources of materials, or any number of possible reasons. This means that the vast number of publications to date have computational chemists comparing their calculations to a single affinity value from the literature and then trying to obtain a precision greater than the likely error in the experimental data.[17−19] Validation and parametrization studies on force fields and affinity predictors should be checked against systems where there are multiple measurements from multiple biophysical techniques that are in agreement. If not, how can we place realistic confidence limits on those calculations? There are a few systems in the literature that have multiple measurements from different biophysical techniques for a series of inhibitors to a given target. One example compares ITC (isothermal titrating calorimetry) and SPR (surface plasmon resonance) by Myszka[21] et al., and another example compares ITC to TSA (ThermoFluor,[22] thermal shift assay) by Matulis[23] et al. More studies such as these are needed to be able to appropriately assess the error or variation in the affinity data. CSAR addressed this issue in our in-house data sets by measuring the affinity by multiple biophysical methods: ITC, ThermoFluor, and Octet RED.[24,25]

This second release of data from CSAR can be found at the following: http://www.csardock.org/MainContent.jsp?page=DataSet.jsp.[26] Below, we outline the selection criteria used to choose data from industrial sources and the subsequent set up of the ligands and proteins for docking and scoring. The full release of the data is described, and the selection of a subset for the 2012 Exercise is explained.

METHODS

Crystallographic Criteria. CSAR has a series of metrics for assessing the quality and suitability of crystal structures for a high quality data set based on a combination of our own experiences[27] and work in the literature. CSAR has adopted the new validation tools used in the PDB.[28] The work of Warren[10] et.al. has influenced a number of our criteria, for example, real space r (RSR) and real space correlation coefficient (RSCC). The CSAR metrics involve comprehensive criteria for assessing the diffraction data, protein structure, and small molecule structure. The criteria that CSAR uses for a CSAR-quality crystal structure is given in Table 1. We use the EDS server[29] in Uppsala to calculate the RSR and RSCC values. The criteria are not blindly applied, and we do a detailed visual inspection of the density and structures to ensure quality.

Table 1. CSAR Criteria for High Quality Crystal Structures

CSAR criteria for diffraction data
- Overall Rmerge ≤ 0.1 (highest resolution bin ≤ 0.4)
- Resolution 2.5 Å or better
- Signal to noise ratio ≥ 2 for 50% or more of reflections in highest resolution bin (preference $I/\sigma I \geq 3$)
- Completeness of data ≥ 90% (highest resolution bin ≥ 50%)
- Redundancy ≥ 2 (low symmetry)
- Redundancy ≥ 3 (high symmetry)
CSAR criteria for protein structure
- Rfree−Rwork ≤ 5%
- Molprobity: protein structure
Poor rotamers < 1%
Ramachandran outliers < 0.2%
Residues with bad bond angles 0%
Residues with bad angles 0%
Clashscore ≤ 5
- Whatcheck
RMS Z scores near 1.0
Torsions
B-factor distribution
Bonds and angles
- Parvarti[46] server: check distribution of anisotropy
Bonds linking atomic displacement parameters (TLS) have correlation coefficient > 0.92
CSAR criteria for structure of small molecule
- No ring puckers
- No eclipsed hydrogens
- Real Space R ≤ 0.2
- Real space correlation coefficient ≥ 0.9
- ≥ 90% of compound atoms in 2Fo−Fc density
- No large unexplained density within 5 Å of compound
- No severe clashes between reduced amino acids
- No symmetry related atoms within 5 Å of compound atom
- No ambiguously fitted compounds: all alternate conformations clearly defined by density
- No more than two alternate conformations of compound may coexist
- Structures created using SMILES input to grade (Global Phasing, Inc.[47]), which uses CCDC Mogul[48] to create known substructures and QM to fill in missing hydrogens and ring torsions. QM performed with imposed Neutron and Mogul restraints

Selection of the CSAR Data Set. The CSAR center has been collecting and curating data from pharmaceutical companies primarily. Some companies choose to compile their own data sets using our criteria, and others welcome our assistance, having us onsite to perform the analysis and selection. Donated crystal data include the refined or partially refined structure coordinate file, structure factor files or .mtz files, and log file of the scaled diffraction data to assess the quality of the diffraction data. These allow the community to verify the quality of the structures. Crystal structures fall into two general categories: CSAR quality and PDB_only. CSAR quality structures have passed all the stringent criteria listed in Table 1 and are included in the online data set for download. CSAR-quality structures are deposited in the PDB. Structures that do not meet the stringent criteria are deposited in the PDB only, hence the designation PDB_only, and are not included in the online set for download. The CSAR center works closely with the company to be sure that all parties are in agreement on the depositions to the PDB and appropriate authorship and acknowledgments are included. We encourage the company to provide unpublished data, including inactive compounds and any activity cliffs. Any form of additional data such as counter assays, measured physical properties, or multiple measurements of affinity are also very beneficial. For internal review and legal release authorization, the company usually treats their data set(s) in the same manner as an external peer reviewed publication, and once approved, an agreement releasing the data for public use is signed.

When a representative from CSAR goes to the company and selects the compounds and finds all the relevant crystal data, a confidentiality agreement is signed in advance. Again, mutual agreement on the selected data is critical. We wish to thank Abbott in particular for generously allowing us to mine their data for submission to CSAR. CSAR canvasses the literature for what targets the company has previously published, including crystal structures, to help reduce the workload on the corporate scientists and make deposition as easy as possible. Once a set of viable targets has been selected, CSAR then extracts the known compounds and affinity data for what has been published on those targets from data sources such as ChemBL[20] and BindingDB.[30] Irrespective of who actually published the data, it will be used at the corporate site to compare the compounds the company has with what is already published so that new unpublished data is gleaned. If a crystal structure is unpublished, but the affinity of the small molecule has been, it is still included as valuable information. Inactive compounds and activity cliffs are also sought for each series. We can, and do, finish structures for companies; often in industry, structures are taken just far enough along to confirm a hypothesis and not fully refined. We can complete the refinement so the company and the community benefit.

Approximately 50 active compounds and 10 inactive ones per series are chosen to provide reasonable coverage of properties and affinities. We often receive much more than 50 compounds per series. Typically, a company has a couple of thousand ligands to choose from. To choose the representative subset, we use recursive partitioning[31] based on the negative log of the affinity versus calculated physical properties: number of hydrogen-bond donors, hydrogen-bond acceptors, number of rotatable bonds, molecular weight, and topological surface area. We split the full set until the log worth has reached its set point (JMP[31] default) or there are no more than five compounds in each individual leaf. This allows us to classify and bin the compounds in a logical fashion using relevant variables. Using this binning in conjunction with visualization of a distribution analysis, we make initial selections in the spreadsheet and then tailor those selections to obtain a more even distribution across the given calculated properties (Figure 1).

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

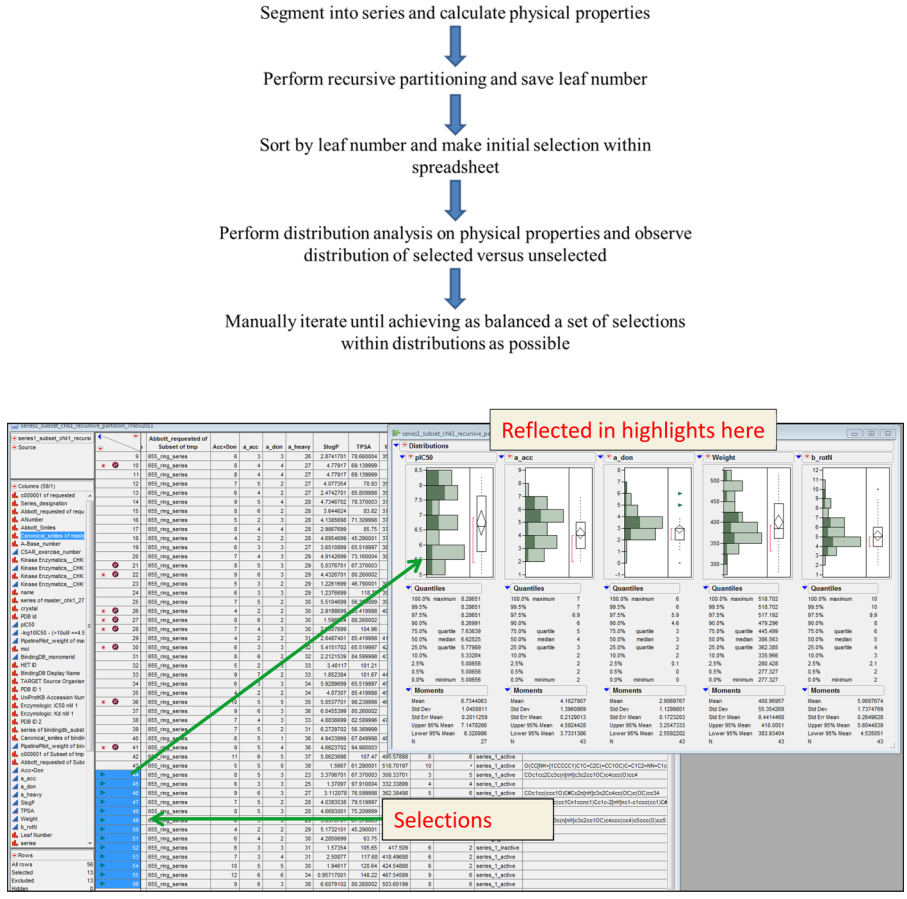


Figure 1. Selection method utilizing recursive partitioning and coupled multiple distribution analysis in JMP[31].

Table 2. 2012 Release Data Set Summary[a]

Target	Total # of Compounds	CSAR Quality Structures	Affinity Range for protein target	Inactive
CDK2	110	19 + 1 apo	0.18mM to 67 nM (K_d) For all compounds	83 (Kinase Known Inactive set)
CDK2-Cyclin A	110	1 + 1 apo	>100μM to 35nM (K_d) For all compounds	83 (Kinase Known Inactive set)
LpXC 2 series	20 (8)	5 (4)	>100 μM to 8 nM (K_d) For all compounds	2 (0)
	12 (8)	0		10 (8)
Chk1 –(Abbott) 3 Series	54 (14)	6 (4)	>10μM to 0.2nM (IC_{50}) For all compounds	9 (3)
	61 (16)	5 (5)		9 (2)
	44 (17)	6 (5)		12 (4)
Urokinase (Abbott)	46 (20)	7 (4)	>50 μM to 0.8nM (K_i) For all compounds	11 (4)
Erk2 (Vertex)	52 (39)	12 (12)	4.8 to 9.4 (pK_i) For all compounds	0
CSAR-NRC (PDB)	343	343	0.05 to 13 (-logK) For all compounds	0

[a]Targets in yellow were used in the 2012 Exercise. Numbers in parentheses indicate the number included in the 2012 Exercise.

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

In-House Projects. The CSAR center has also pursued our own in-house effort to compile data sets. Octet RED[25] and nanoITC[32] are preferred for obtaining affinity, kinetic, and thermodynamic information. We also use a ThermoFluor system (thermal stability assay) to measure affinity. By having three different biophysical techniques used by the same personnel at the same lab maintaining the conditions as similarly as possible, the variance should be low, and the affinities produced should be as accurate as possible. Measurements are made in triplicate or more for each K_d value, and we have gone as high as 16 to obtain an adequate assessment of the variance. Ligands have been synthesized by the University of Michigan Vahlteich Medicinal Chemistry Core[33] and WuxiApptec.[34] WuxiApptec also measured the physical properties of all the compounds used in our internal efforts. The properties measured for the compounds are thermodynamic solubility, logD (pH 7.4), logP (pH 11) and pK_a values. Protein production and crystallography are done by CSAR staff.

Protein−Ligand Complexes and Docking Decoys. As noted above, all CSAR-quality crystal structures are set up for docking and scoring. We also chose to generate a diverse set of decoy binding poses using DOCK[35−38] (version 6.5) for all 58 ligands. The negative image of a molecular surface was generated using the SPHGEN[39] utility, and the binding site was represented by all the spheres within 12 Å of any ligand atom (in five complexes, it was increased to 15 Å due to large ligand size). The potential grid was precalculated using the GRID program with a grid spacing of 0.3 Å and with additional 3 Å boundaries in each direction of the grid. Flexible ligand docking was applied using the anchor-and-grow algorithm.[36] The default set of parameters were used, and DOCK[35−38] poses were clustered with a 1.0 Å cutoff. In order to provide sufficient sampling, the maximum number of orientations was set to 2000; however, varying number of poses (∼200−1800) were obtained for different systems.

A diverse set of 200 poses (in terms of RMSD) was selected using the ranking obtained from the "Diverse Subset" utility in MOE 2011.10.[40] Initially, poses were visually inspected, and poses outside the binding side or scored very low (>0 kcal/mol, if total number of poses were at least 200) were removed. For analysis, the symmetry-corrected RMSD between ligand poses was calculated using the SVL script provided by support scientists at the Chemical Computing Group.[40] At least one near-native pose (RMSD < 1.0 Å) was found for all systems except for six Chk1 cases (ring−urea−ring series). The near-native poses for these six cases were obtained by rigid ligand docking using DOCK[35−38] and are included in their decoy sets. To make the clear distinction between the right and the wrong decoy poses, all the poses with RMSD < 2 Å were discarded except for one near-native pose (RMSD < 1.0 Å).

RESULTS AND DISCUSSION

Table 2 summarizes the data set we have just released. These targets have, in general, multiple series with crystal structures and inactive compounds for each series. Inactive compounds are rarely published in the literature, and the CSAR center makes a point to obtain inactive compounds for every series that it can. The download site contains all of the details of the available protocols, crystal structures, SMILES strings, thermodynamic values where possible, affinities, and error estimates (when possible), along with the measured physical properties also when available. The PDB codes for all the structures and the designations (i.e., CSAR quality) are also given.

Chk1 from Abbott has 106 crystal structures in the PDB. Of these, CSAR has deposited 30 new structures and one rerefined structure from the Abbott Chk1 submission (19 CSAR-quality structures). To place this in perspective, the CSAR-NRC data set has all the high quality crystal structures from the PDB up to 2008, and it contained only six PDB ids for the EC 2.7.11.1, which includes Chk1. There are eight structures for a new 7−6−7 ring system core that had no examples in the PDB previously, and 12 new structures augmenting a second series (6−5−5 ring system) to two existing structures for this second series. In a third series (ring−urea−ring series), six new structures have been added to the PDB to augment the four existing structures.

Urokinase has 301 structures in the PDB with four to be released. The CSAR-NRC HiQ data set contained only 10 PDB ids for EC 3.4.21.73, which includes urokinase. CSAR has added nine new structures (seven CSAR quality) and rerefined three additional ones from the Abbott submission. Notably, in the new structures added and in the three rerefined ones, a succinic acid molecule in close proximity to the ligand was identified. These were not identified in the existing structures, but most likely, they were present in all 10 prior structures.

Erk2 has 54 structures in the PDB, and of these, 42 have small molecule inhibitors. Very few met CSAR-quality metrics. In fact, the CSAR-NRC HiQ data set contained only three PDB ids for EC 2.7.11.24, which includes erk2. There are 14 CSAR-deposited structures from the Vertex submission (12 CSAR quality) filling out a diverse combination of five and six member rings or 5−6 ring systems in the series.

LpxC from *Pseudomonas aeruginosa*, a CSAR internal project, has four crystal structures in the PDB. The CSAR-NRC HiQ data set contained no LpxC crystal structures. CSAR has submitted five additional CSAR-quality crystal structures from our own work. The compounds were synthesized at the Vahlteich Medicinal Chemistry Core[33] at the University of Michigan. See Supporting Information for the protocols the CSAR center used for the affinity assays.

CDK2 and CDK2-cyclinA, two CSAR internal projects, were used as an example of a kinase (a popular target class) and as a project to assess the effect of the coactivator cyclinA. The CSAR-NRC data set contained only two PDB ids for EC 2.7.11.22, which includes CDK2. Using these two systems, we have added 21 crystal structures to the PDB (16 CSAR quality) and have kinetic data from three sources (Thermofluor, ITC, and Octet RED). We also created a tethered cyclinA-CDK2 complex, so that we were able to get the kinetic data for ligand binding without the protein−protein binding skewing the data from the Octet RED. See Supporting Information for the protocols the CSAR Center used for the affinity assays.

We created a "Known Kinase Inactives" set. The set is a library of 85 compounds purchased from Chembridge[41] that has passed a pharmacophore search for a general ATP−kinase binding site. CSAR chose the compounds that were within 77−90% similar to known kinase inhibitors (found in PubChem[42] at the time of order). All compounds are chemically plausible as kinase inhibitors and experimentally tested to confirm inactivity with CDK2 and CDK2-cyclinA. Our protocol is to test this set against every kinase (or ATP-binding site) that CSAR works on internally. This provides us with a consistent set of inactives across similar binding sites. If any compound is active against a future target, of course, it will be included as an active and

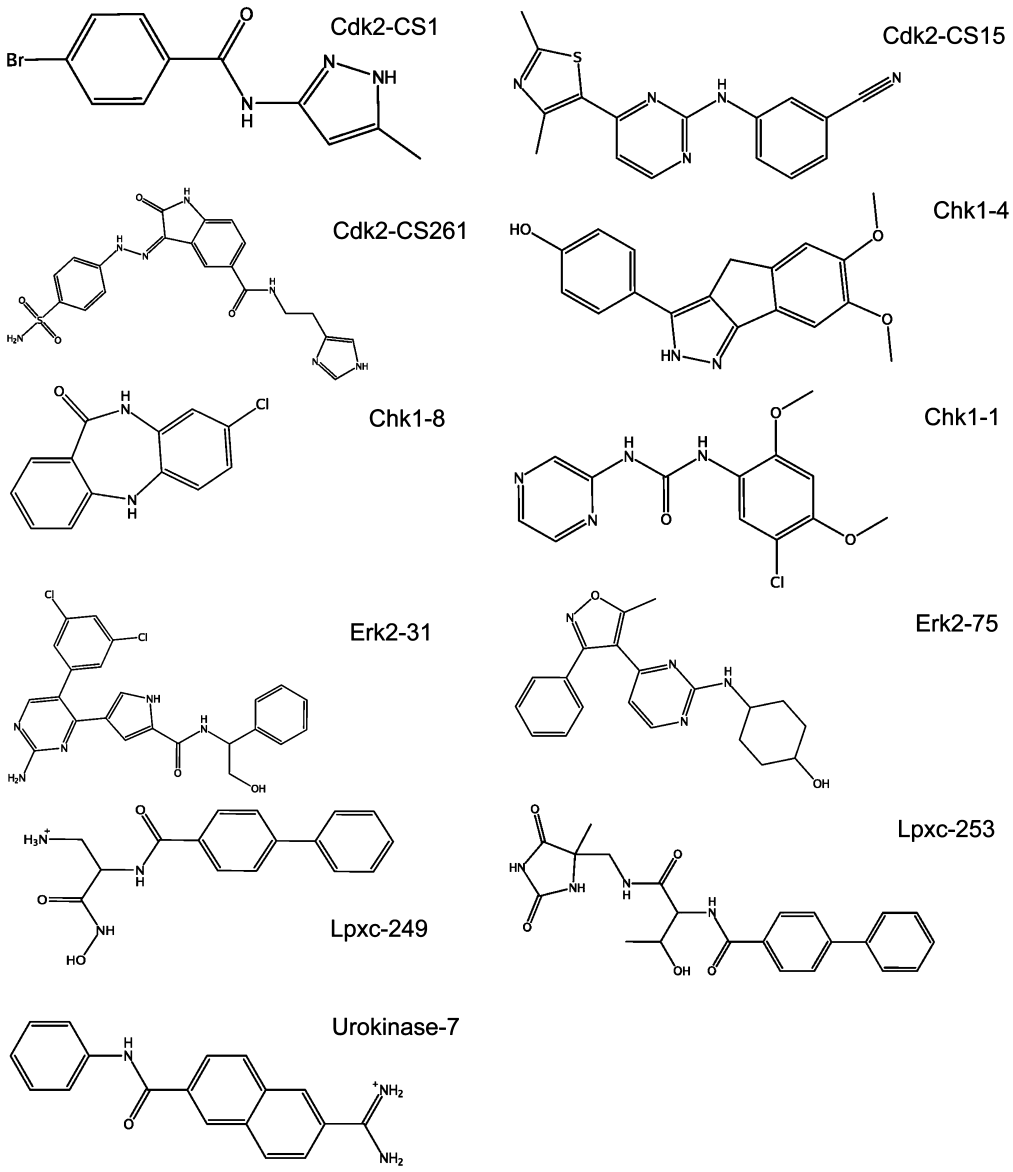


Figure 2. Representative ligands in the 2012 release data set.

removed from the inactives list for that target. The physical properties for the 85 compounds have been measured (logD, solubility, etc.) and are available in the data set download. This set has only been tested against CSAR in-house projects.

In total, the current release of CSAR data comprises 647 compounds, 6 targets, and 82 crystal structures deposited in the PDB (59 are CSAR quality). There are inactive compounds for each series except for erk2, and we are working with Vertex to obtain some. For the in-house projects, we have measured pK_a, solubility, thermodynamic data, logD, and logP data and measured affinity using three different biophysical methods. See Figure 2 for a representative set of ligands contained in the data set.

Docking and Scoring Structures, Including Decoys. The data set also contains the compounds and crystal structures set up in a consistent fashion, including the protonation and tautomeric state, ready for docking and scoring. We have created decoy sets of ligand poses for the six protein targets (58 complexes) in the current CSAR benchmark. This provides test sets for scoring functions by decoupling the scoring problem from the sampling problem. Decoy sets are also available for all the structures of the CSAR-NRC HiQ benchmark release,[18] which were provided by Prof. Xiaoqin Zou's group.[43] In the decoy sets for the second benchmark release, we have ensured that the decoys fully cover the binding site, are diverse in their poses, and include exactly one near-native pose. There is clearly diversity within the poses, yet clearly one correct answer.

Figure 3 shows the distribution of RMSDs between the decoy poses and the native pose for each of the six targets. A normal distribution with RMSD ranging from 1 to 22 Å and peaks of ∼10−15 Å can be seen for the different targets. As per the design, only one near-native pose (RMSD < 1.0 Å) and no pose in the 1−2 Å range is evident in the distribution plots. This makes a clear distinction between the near-native pose and other decoy poses in evaluating scoring functions. We acknowledge that the set may be biased by the force field used, and users should determine whether their methods will be significantly affected. Minimization of the ligand with a rigid protein may be necessary in some cases.

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

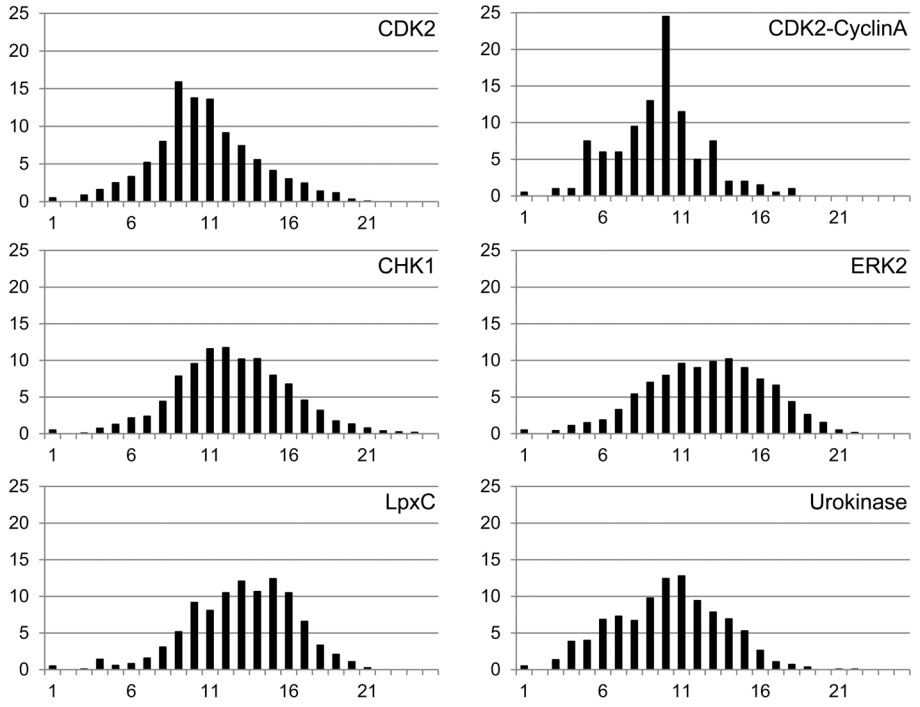


Figure 3. Percentage frequencies of RMSDs (Å) between decoy poses and the native crystal pose for the different targets. The frequencies are based on all the structures available for each target.

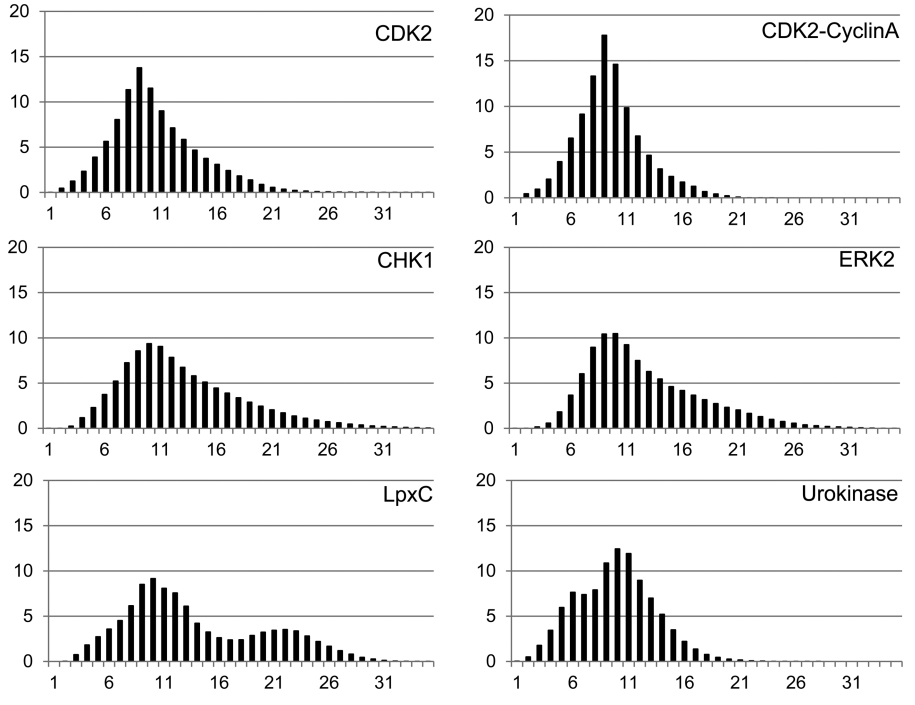


Figure 4. Percentage frequencies of RMSDs (Å) between decoy poses themselves for the different targets. The frequencies are based on all the structures available for each target.

Figure 3 provides the distribution of RMSD between the crystal poses and the decoys, but Figure 4 gives the distribution of RMSDs between the decoy poses themselves. Overall, positively skewed normal distributions are seen for the different targets, with RMSD values as large as 35 Å, peaking at ∼9−10 Å. A very small number (on average ∼0.25%) of RMSD values lies in the 1 and 2 Å bins, which indicates that the decoy sets are distinct from each other just like they are distinct from the crystal pose. The differences in the overall distribution in different targets reflect the differences in the shape of the binding pockets. For example, LpxC shows a bimodal distribution with a major mode of ∼10 Å and a minor mode of ∼22 Å. This is caused by the binding pocket having two distinct openings. A representative structure from each target overlaid with the native and decoy poses is shown in Figure 5,

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

which illustrates the comprehensive coverage of the binding site by the decoy sets for the different targets.

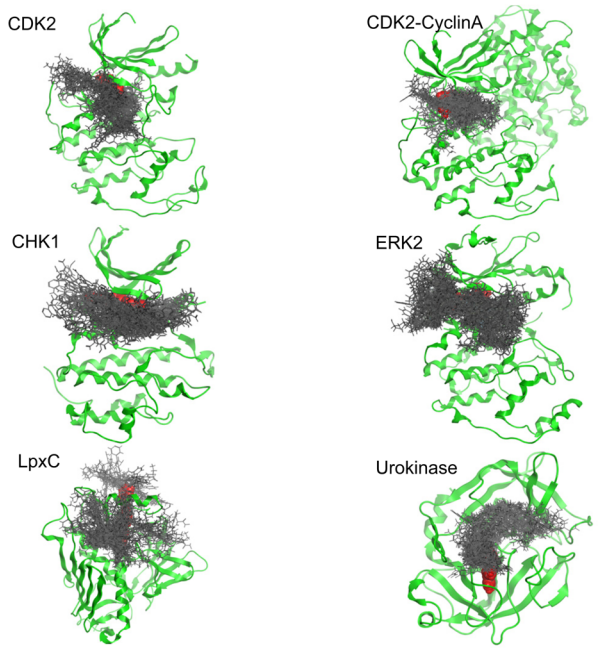


Figure 5. Representative of the different targets (green) with the native bound pose (red) and the 200 decoy poses (gray). The representative protein–ligand complexes are CDK2-CS12, CDK2-CyclinA-CS260, CHK1-70, ERK2-000075, LpxC-CS252, and Urokinase-15 (second term is the ligand number in the data set).

What Is the Affinity? One of the goals of the CSAR center is to provide multiple measurements from multiple biophysical methods in order to assess the variation in affinity data. The vast majority of affinity values in the literature are from a single measurement, which makes it difficult to assess the accuracy of a docking and scoring method. There are a large number of possible assays and conditions that the system is amenable to, and it is best to measure affinity under many different conditions in order to know the true value and its variance. If the affinity value is consistent between assay methods, an important source of uncertainty has been removed. In the literature, data like this is frequently given when scientists develop new assay methods and equipment, publishing studies on comparing the new technique to established methods. The hsp90 TSA study by Matulis[23] and carbonic anhydrase SPR project by Myszka,[21] mentioned earlier in the manuscript, have shown good correspondence with ITC measurements. While these two studies showed good agreement, some do not. A carbonic anhydrase study from Jecklin, et al.[44] showed that most of the compounds K_d values correlated well by the three methods, but a few of the compounds did not correlate well at all. This behavior seemed to be compound specific.

CSAR is using three biophysical methods allowing comparison of ThermoFluor, OctetRed, and ITC data to examine the variance across the methods. CDK2 is CSAR's in-house system with the most data to date. Each individual data point, for each method, is an average of a minimum of three K_d measurements per method. In general, our affinities have a relatively low standard error of measurement. Additionally, each experiment has been performed to the best of our ability in conditions as identical as possible. Even given all the effort to keep the individual variance in the data low, the correlation of the K_d values from ThermoFluor (TSA) with those from the Octet RED has an R^2 of about 0.4. While the correlation is positive, we would have liked to see better agreement in the absolute values of the measured affinity. Often in the literature, when an affinity value is double checked by another technique, that affinity value is compared to one measured by ITC. If we choose the K_d values from our ITC measurements as our "gold standard" for comparison, the Spearman ρ for ITC to TSA is 0.77 and for ITC to Octet RED is 0.83. For comparison, the Spearman ρ for TSA to Octet RED is 0.59. Spearman ρ is an appropriate statistic for examining relative ranking, as opposed to the absolute value correlation discussed previously. The *relative rankings* of both ThermoFluor to ITC and Octet RED to ITC are reasonably correlated. See Figure 6 for details of the multivariate analysis of the data. An unfortunate limitation is few ITC data to compare to other methods. Though it is considered a gold standard, ITC is not amenable to all targets and ligands. ITC has a narrower affinity range than TSA, SPR, or Octet RED, and it is more sensitive to solubility problems associated with typical drug-like compounds. This limits the number of affinities that can be determined by the method. In our case, the number of ITC measurements for CDK2 is only six, and if one wanted to use the most appropriate K_d values from another method, it would appear that those from the Octet RED would be best based on Spearman ρ value.

In the CDK2 study, we are using techniques that have fundamental differences in how they measure affinity. Octet RED uses immobilized protein, and ThermoFluor utilizes a change in temperature. It appears that in some instances (apparently compound/target specific), the measured affinities do not correlate well. As part of this analysis, we have looked for any possible explanation that could account for the differences in measured physical properties, but we have not identified any factors that would explain the affinity differences. It may be prudent to validate and parametrize docking methods using only affinity data that correlate well between multiple biophysical methods.

The CDK2-cyclinA data we have obtained would indicate that most but not all compounds bind even tighter than with CDK2 alone. Interestingly, we were unable to get a crystal structure for one of the compounds (CS17) bound to CDK2, despite a pK_d of 5.1 in ThermoFluor and 7.3 in Octet RED (the largest variance in pK_d for all the compounds). While it may be solubility related, CS17's solubility was no worse than other compounds for which we were able to get crystal structures.

This indicates that there can be considerable variation in the data between methods, even when great care is taken to do the measurements as similarly as possible. Looking at the SEM values, each individual data point is low. On the basis of comparison with the other methods we employed, Octet RED tends to give a slightly lower K_d that the ThermoFluor or ITC for this set of compounds and for this target. All of the affinity data and SEM data is in the spreadsheets contained in the download set on the CSAR Web site (www.csardock.org). We are working on hsp90 and currently have ITC measurements for 18 compounds and 27 compounds for which we are obtaining ThermoFluor and Octet RED K_d data. This will provide another system for comparison allowing us to identify if the variance may be target dependent or dependent on the methods employed. If the variance is target and/or compound dependent, as may well be the case given the results of the 2012 exercise by Damm-Ganamet, et.al.[45], then it may be wise to

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842–1852

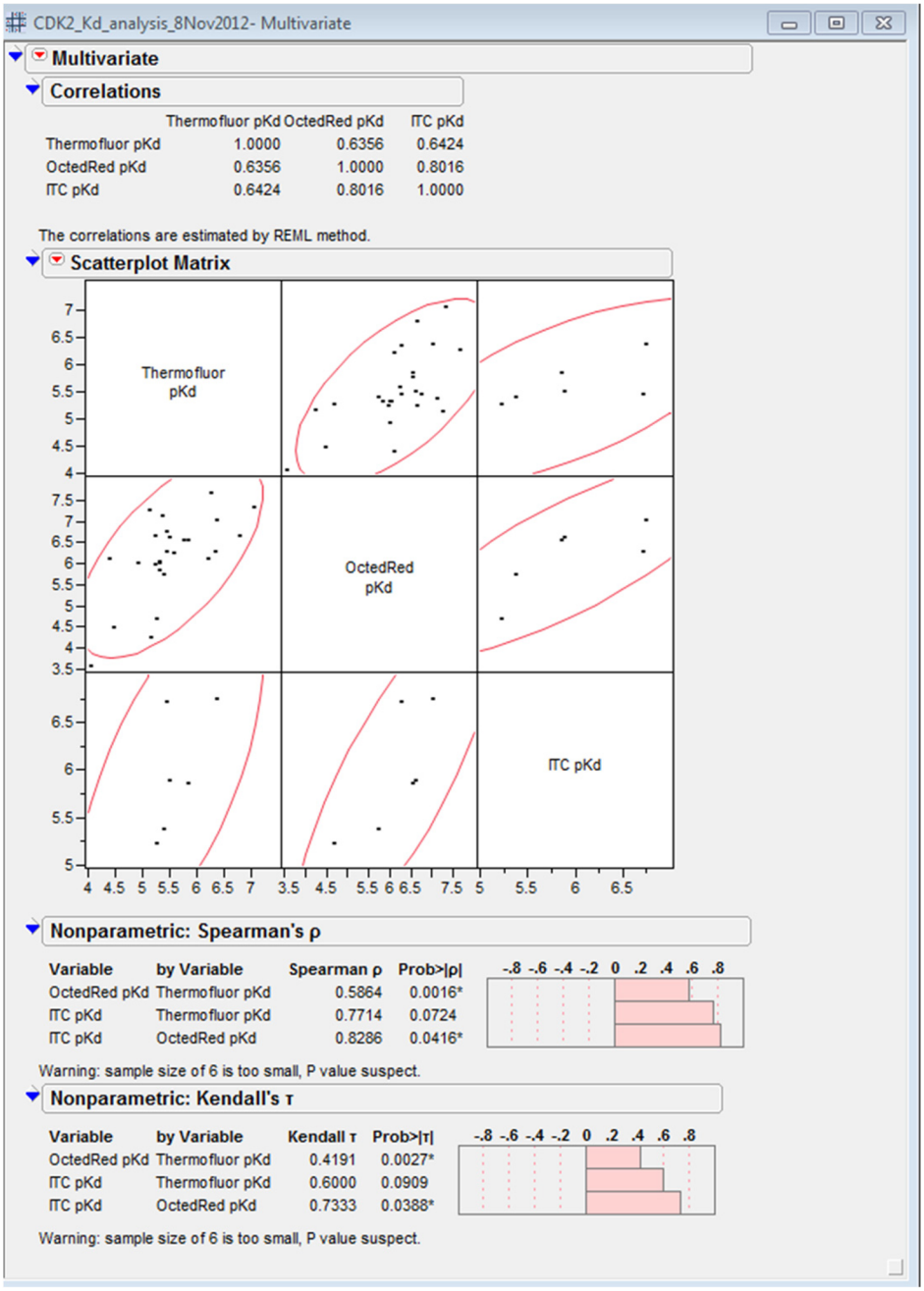


Figure 6. Multivariate analysis of the CDK2 pK_d data in JMP:[31] r, ρ, τ.

parametrize and validate with the targets that have the least variation in affinity data by multiple methods. Experiments are considered to be the gold standard, and computational methods can only be as good as the underlying data. Understanding and working within the variance of the data is critical to parametrizing and evaluating the performance of our methods.

Selection of Exercise Data Set. The data set used for the 2012 exercise was selected from the larger data set described above. A subset of 122 ligands (total for all four targets) was created to make the docking and scoring possible within the time constraints given to the participants. The release of the full data set was held until after the exercise was complete. The exercise had two parts: docking evaluation and affinity prediction. The chosen ligands, both active and inactive, were given to the participants as a file of SMILES strings. The participants were also told which crystal structure from the PDB would be the best to use. The data selected for the docking phase was comprised of all the unpublished CSAR-quality crystal structures. Fifteen of these unpublished crystal structures had a published affinity and were therefore not used in the analysis of the affinity prediction.

The subset employed in the affinity prediction, used compounds with unpublished affinity, were selected to cover as wide a range of affinity as possible. In one set, erk2, we did not have any examples of inactive compounds, so none were included. We targeted approximately 10 compounds per series per target as a reasonable number to predict based on the time frames involved for the exercise and on feedback from the community. The compounds were chosen in a similar fashion as the larger data set: utilizing recursive partitioning with the

dx.doi.org/10.1021/ci4000486 | J. Chem. Inf. Model. 2013, 53, 1842−1852

same properties and manual selection visualized in a distribution analysis (Figure 1). See the Supporting Information for the 2012 exercise subset.

The results of the CSAR center's analysis of the 2012 exercise is presented in an accompanying paper by Damm-Ganamet, et.al.[45] Twenty participants worldwide used this data set and were asked to submit multiple methods in order to test a hypothesis of their choosing. Some participants present their work in this same issue.

CONCLUSION

The CSAR center has released its second major data set on 6 protein targets with 647 compounds and 82 crystal structures comprised of mostly industrial data. A representative set of compounds in the data set, one from each series, is depicted in Figure 2. The in-house systems included in this release have multiple K_d measurements from multiple methods along with measured physical properties (solubility, logD, logP, and pK_a) of the compounds. Additionally, there is a "Known Kinase Inactives" data set and an extensive docking decoy set for each of the 58 CSAR-quality crystal structures. In the docking decoy sets for this second benchmark release, we have ensured that the decoys fully cover the binding site, are diverse in their poses, and include some near-native poses (http://www.csardock.org/MainContent.jsp?page=DataSet.jsp).[26] For a future release, we are already processing data on five new targets from pharma. In addition, we are working on three targets in-house (hsp90, urokinase, and chk1) to provide affinities from multiple biophysical methods. The measured physical properties for the ligands will also be available. We are continuing to work with pharma colleagues for more industrial data sets.

ASSOCIATED CONTENT

Supporting Information

Protocols for the assays (ThermoFluor, ITC, and Octet RED) to obtain the K_d information for CDK2, CDK2-cyclinA, and LpxC projects as well as for the protein purification. The crystallographic protocol used for all CSAR generated and refined crystal structures as well as any submitted crystal structure assessment. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: carlsonh@umich.edu (H.A.C); jbdunbar@umich.edu (J.B.D.). Phone: (734)615-6841 (H.A.C.); (734)615-9092 (J.B.D.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The CSAR Center is funded by the National Institute of General Medical Sciences (UO1 GM086873). We thank Chemical Computing Group and OpenEye Software for the generous donation of the use of their software. We thank our colleagues at the Vahlteich Center for compounds and William Clay Brown for protein expression. Use of the Advanced Photon Source, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the U.S. DOE under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817). We thank the staff of LS-CAT, namely David Smith, for aiding in diffraction screening of crystals for in-house data sets.

REFERENCES

(1) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(2) Corbeil, C.; Williams, C.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−786.

(3) Blaney, J. A very short history of structure-based design: How did we get here and where do we need to go? *J. Comput.-Aided Mol. Des.* **2012**, *26*, 13−14.

(4) Green, D.; Leach, A.; Head, M. Computer-aided molecular design under the SWOTlight. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 51−56.

(5) Segall, M. Can we really do computer-aided drug design? *J. Comput.-Aided Mol. Des.* **2012**, *26*, 121−124.

(6) Woltosz, W. S. If we designed airplanes like we design drugs... *J. Comput.-Aided Mol. Des.* **2011**, *26*, 159−163.

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

(8) Welcome to the Worldwide Protein Data Bank. http://www.wwpdb.org/ (accessed November 29, 2012).

(9) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(10) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential considerations for using protein−ligand structures in drug discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.

(11) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein−ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.

(12) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An Extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.

(13) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(14) Favia, A. D.; Bottegoni, G.; Nobeli, I.; Bisignano, P.; Cavalli, A. SERAPhiC: A benchmark for in silico fragment-based drug design. *J. Chem. Inf. Model.* **2011**, *51*, 2882−2896.

(15) Stouch, T. R. The errors of our ways: Taking account of error in computer-aided drug design to build confidence intervals for our next 25 years. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 125−134.

(16) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the prediction of absolute solvation free energies using the next generation OPLS force field. *J. Chem. Theory Comput.* **2012**, *8*, 2553−2558.

(17) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy skepticism: Assessing realistic model performance. *Drug Discovery Today* **2009**, *14*, 420−427.

(18) Smith, R. D.; Dunbar, J. B.; Ung, P. M.-U.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115−2131.

(19) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public *Ki* data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(20) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.

(21) Day, Y. S. N.; Baird, C. L.; Rich, R. L.; Myszka, D. G. Direct comparison of binding equilibrium, thermodynamic, and rate constants determined by surface- and solution-based biophysical methods. *Protein Sci.* **2002**, *11*, 1017−1025.
(22) Pantoliano, M. W.; Petrella, E. C.; Kwasnoski, J. D.; Lobanov, V. S.; Myslik, J.; Graf, E.; Carver, T.; Asel, E.; Springer, B. A.; Lane, P.; Salemme, F. R. High-density miniaturized thermal shift assays as a general strategy for drug discovery. *J. Biomol. Screening* **2001**, *6*, 429−440.
(23) Kazlauskas, E.; Petrikaitė, V.; Michailovienė, V.; Revuckienė, J.; Matulienė, J.; Grinius, L.; Matulis, D. Thermodynamics of aryl-dihydroxyphenyl-thiadiazole binding to human Hsp90. *PLoS ONE* **2012**, *7*, e36899.
(24) Abdiche, Y.; Malashock, D.; Pinkerton, A.; Pons, J. Determining kinetics and affinities of protein interactions using a parallel real-time label-free biosensor, the Octet. *Anal. Biochem.* **2008**, *377*, 209−217.
(25) Octet RED96 System. ForteBio. http://www.fortebio.com/octet_RED96.html (accessed November 27, 2012).
(26) 2012 Datasets. CSARdock.org. http://www.csardock.org/MainContent.jsp?page=DataSet.jsp (accessed December 19, 2012).
(27) Dunbar, J. B.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Selection of the protein−ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036−2046.
(28) Read, R. J.; Adams, P. D.; Arendall, W. B.; Brunger, A. T.; Emsley, P.; Joosten, R. P.; Kleywegt, G. J.; Krissinel, E. B.; Lütteke, T.; Otwinowski, Z.; Perrakis, A.; Richardson, J. S.; Sheffler, W. H.; Smith, J. L.; Tickle, I. J.; Vriend, G.; Zwart, P. H. A new generation of crystallographic validation tools for the protein data bank. *Structure* **2011**, *19*, 1395−1412.
(29) Kleywegt, G. J.; Harris, M. R.; Zou, J.; Taylor, T. C.; Wählby, A.; Jones, T. A. The Uppsala electron-density server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240−2249.
(30) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(31) JMP Software, Data Analysis, Statistics, Six Sigma, DOE (Version 8). JMP. http://www.jmp.com/ (accessed November 27, 2012).
(32) Nano ITC. TA Instruments. http://www.tainstruments.com/main.aspx?siteid=11&id=263&n=3 (accessed November 29, 2012).
(33) Hans W. Vahlteich Medicinal Chemistry Core. http://sitemaker.umich.edu/mccsl/home (accessed November 27, 2012).
(34) Integrated R & D Services. WuXi AppTec. http://www.wuxiapptec.com/ (accessed November 27, 2012).
(35) Shoichet, B. K.; Kuntz, I. D.; Bodian, D. L. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397.
(36) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(37) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601−619.
(38) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: Combining techniques to model RNA−small molecule complexes. *RNA* **2009**, *15*, 1219−1230.
(39) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722−729.
(40) Chemical Computing Group. http://www.chemcomp.com/ (accessed December 4, 2012).
(41) Screening Libraries. Targeted & Focused Libraries. ChemBridge. http://www.chembridge.com/screening_libraries/targeted_libraries/?PHPSESSID=62cd1ffe32f7ad197c98e923c1006053 (accessed November 29, 2012).
(42) The PubChem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed November 30, 2012).
(43) Huang, S.-Y.; Zou, X. Construction and test of ligand decoy sets using MDock: Community structure−activity resource benchmarks for binding mode prediction. *J. Chem. Inf. Model.* **2011**, *51*, 2107−2114.
(44) Jecklin, M. C.; Schauer, S.; Dumelin, C. E.; Zenobi, R. Label-free determination of protein−ligand binding constants using mass spectrometry and validation using surface plasmon resonance and isothermal titration calorimetry. *J. Mol. Recognit.* **2009**, *22*, 319−329.
(45) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B., Jr.; Stuckey, J. A.; Carlson, H. A. CSAR Benchmark Exercise 2011−2012: Evaluation of results from docking and relative ranking of blinded congeneric series. *J. Chem. Inf. Model.* **2013**, DOI: 10.1021/ci400025f.
(46) PARVATI: Protein Anisotropic Refinement Validation and Analysis. http://skuld.bmsc.washington.edu/parvati/ (accessed December 4, 2012).
(47) Global Phasing Limited. http://www.globalphasing.com/ (accessed December 4, 2012).
(48) Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; Orpen, A. G. Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2133−2144.