Published in final edited form as:
J Chem Inf Model. 2016 June 27; 56(6): 951–954. doi:10.1021/acs.jcim.6b00182.

Lessons Learned over Four Benchmark Exercises from the Community Structure-Activity Resource

Heather A. Carlson
Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, 428 Church St., Ann Arbor, Michigan 48109-1065, United States

Abstract

Preparing datasets and analyzing the results is difficult and time-consuming, and I hope the points raised here will help other scientists avoid some of the thorny issues we wrestled with.

Introduction

People have asked me what lessons we have learned from running four CSAR Benchmark Exercises.[1–6] Some lessons simply taught me about human nature: 1) few will actually turn in their data in the requested format, 2) there will always be people who need an extension of the deadline, and 3) some people will not like the answers. Also, despite being common knowledge, I had to learn the hard way that you really can't make everybody happy. More importantly, the Exercises were an opportunity to think deeply about our science, question our basic assumptions, and engage colleagues in thoughtful discussion.

Lesson 1: Good crystal structures are hard to find

I must thank Greg Warren of OpenEye for teaching me this valuable lesson! In our first Exercise,[1,2] we focused on identifying excellent crystal structures with really pristine electron density in the binding site (see Figure 1). One of the most important features of a good structure is that the ligand have well resolved density and a real-space correlation coefficient (RSCC) of ≥0.9.[7] The structures were also required to have no contacts to the ligand from crystal additives or symmetry packing, so that any artificial deformation of the ligand coordinates could be avoided. If the coordinates of the ligand are poorly resolved or artificially deformed, any disagreement between the docked pose and the crystal coordinates is meaningless. Regardless, that meaningless disagreement would result in a higher RMSD and an unfair, negative assessment of the docking method. We started with the entire Protein Data Bank (PDB)[8] from 2008 and pared it down from 47,132 structures to 342! Less than 1% of the PDB structures met our "HiQ" definition. Recently, our HiQ criteria have been adopted by PDBbind for their "core" set,[9] and we are pleased this same rigor is being adopted by other major resources for docking and scoring.

Corresponding author: carlsonh@umich.edu.

Lesson 2: Several metrics are needed for assessing docking and scoring

Now that we have good crystal structures to use, we must do proper assessments of docking methods. The most common metric to assess docked poses is root mean squared deviation (RMSD) of ligands, where one compares the heavy atom positions from the pose to the crystallographic coordinates of the bound ligand. Corrections must be made if there is any symmetry in the ligand or the pocket, and RMSD ≤ 2.0 Å is the accepted definition of a correct pose. However, RMSD is not necessarily the best metric when cross-docking ligands or when trying to include protein flexibility in docking because choosing a frame of reference is subjective. Structural alignments can be based on the protein backbone or on the binding-site residues. If aligning the binding sites, one must choose which residues to include. For our previous study,[4] we chose to superimpose protein structures by their backbones, using our Gaussian-weighted method (wRMSD,[10] which is available in that paper's supplemental information and in the MOE software package[11] as the "Gaussian-weighted" option for superimposing structures). Our method emphasizes the most agreement between two structures, whereas standard RMSD places the greatest mathematical emphasis on the positions that are most different between the structures. With wRMSD, the frame of reference is dictated by the most common core structure of the protein. These were our choices for calculating RMSD of submitted ligand poses in our 2012 Exercise, but different test systems might call for alternate choices.

It is important to remember that most crystal structures only provide a single, static snapshot of a protein-ligand complex. If an NMR structure exists for the same complex, each structure in the NMR ensemble will have a different RMSD when compared to the crystal structure. This is true even if the same contacts are maintained between the ligand and the protein in each structure of the ensemble. Clearly, all of those NMR structures are correct answers. To account for this perspective, we also measured protein-ligand contacts when evaluating docked poses as an alternative metric to RMSD.[4] There is still some subjectivity when choosing which residues to include and what cutoffs to use when measuring the contacts, but the choices are relatively straightforward. Corrections are needed for ligand symmetry, and "equivalent" contacts to side chains must be counted as correct (contacts to CD1 are equivalent to contacts to CD2 in Leu, OE1 is equivalent to OE2 in Glu, NH1 is equivalent to NH2 in Arg, etc). We examined all protein atoms within 4.0 Å of the ligand's non-hydrogen atoms. The total count of those protein atoms is a measure of "general packing" that captures both tight and loose van der Waals (vdw) interactions provided to the ligand. Our cutoff for hydrogen-bonding and electrostatic interactions was the standard value of ≤ 3.5 Å (first-row atoms N, O, and F), but the cutoff to larger atoms (S, Br, Cl, etc) was slightly longer at ≤ 3.8 Å. For ligands coordinated to metal ions, we recommend ≤ 2.8 Å as a contact cutoff, but the value will likely be system dependent. The contacts for docked poses can then be compared to the native contacts in a crystal structure or NMR structure. A benefit of counting protein-ligand contacts are the insights they give about the limitations of the pair-wise potentials that underlie the docking method. For example, it is possible to learn whether vdw contacts are sacrificed to make additional hetero-hetero contacts with hydrogen bonds. If a method's poses have systematically high total-packing counts, it may indicate a need to introduce desolvation terms. Systematic miscounts for a particular functional group point to poor parameterization. This information is more difficult to find when focusing on RMSDs.

When comparing experimental affinities to scoring/ranking ligands, we used Pearson R, R^2, Spearman ρ, and Kendall τ.[1–6] Those four values cover the most common parametric and non-parametric tests for relative ranking. When assessing a method's ability to discriminate active from inactive compounds, receiver operator curves (ROC plots) were used and quantified by area under the curve (AUC). Enrichment rates are also useful to quantify a method's ability to weed out inactive compounds from a dataset. At this time, computational methods for docking and scoring are good at identifying actives over inactives, but relative ranking of actives is very difficult.

Lesson 3: Embrace statistics, error bars, and confidence intervals

I am thoroughly convinced that the limited progress in scoring/ranking comes from ignoring statistics in our studies. There is no excuse for publishing results without some measure of error. Even if a scientist does not know how to analytically calculate standard deviations or confidence intervals (ci), these can be estimated through bootstrapping. Bootstrapping is used to randomly sample your dataset with replacement, meaning the same data point can be chosen more than once (sampling without replacement is jack-knifing which gives smaller estimates of error). Scientists may choose different sample sizes, but in general, one repeatedly chooses a random subset of data and re-calculates their chosen metric (eg, AUC). If we have N ligands scored by a ranking method, our procedure to estimate error for AUC is to randomly sample 90% of those N scores (with replacement) and repeat this 10,000 times, which gives a distribution of 10,000 AUC values. That distribution provides mean, median, standard deviation, and 95% ci of AUC for any scoring/ranking method. For clarity, it should be noted that the 95% ci is the range of values from 2.5% to 97.5% of the distribution, which leaves out the extreme 5% of AUC values divided between both tails of the distribution. The 95% ci is the estimate of the range of AUC values likely – 95% of the time – when the method is applied to a new data set. Think of how wildly different the results frequently are for a method applied to two different systems!

In our first exercise,[1,2] the statistics showed that the great majority of methods were equivalent. With so many performing basically the same, I specifically chose not to name which methods gave which results (results were labeled as code 1, code 2, etc). Many people strongly disagreed with that choice; they wanted a ranking of different software packages. They argued that even if methods were equivalent, it would be possible to learn something from the trends. That is not true. The great majority of methods – 13 of 17 submissions plus the two "null" metrics – had Spearman ρ ranging 0.64–0.53 with 95% ci spanning roughly ±0.07. When the confidence intervals overlap so significantly, it means that doing the same test using different protein-ligand complexes will scramble the ordering of the methods. The trends people wanted were meaningless, despite having a dataset of almost 350 data points. However, two methods with $\rho > 0.7$ were equivalent to one another and had statistically significantly higher ρ than the rest. Of course, those successes had important insights to share, and they were discussed in full detail by the submitters in their own papers for our first special issue.[12,13]

The statistics underpinning our evaluation metrics show that datasets need to be at least 1000 ligands with affinities, which are not available at this time.[14] *Frankly, most datasets used in our field are so small that we have been chasing random noise.* Large datasets produce 95% ci that are small enough to clearly distinguish what improvements move a ranking method from below average ($\rho = 0.4$) to acceptable ($\rho = 0.5$) to improved ($\rho \geq 0.6$). The submissions we have received over the years show that this is the "critical" range for improving the majority of current methods across the whole field. Unfortunately, those ρ values are so low that we need a large number of ligand affinities to get statistically significant results. Suppose you wanted to introduce a new term to your ranking method. You need a dataset of 864 to show that the new version's $\rho = 0.65$ is better than the old approach's $\rho = 0.55$ because their 95% ci do not overlap.[14] To say $\rho = 0.65$ is better than $\rho = 0.6$, you need 2974 affinities.[14] No datasets are anywhere close to that size, which means there is less likelihood that new terms are real improvements in our methods. The overwhelming majority of scoring studies use a few hundred complexes and compare "top" methods that differ by $\Delta\rho = 0.05–0.15$. There is no statistical significance to those differences with so few complexes. I challenge the reader to re-examine their favorite scoring papers.

Lesson 4: Making a good dataset is a difficult multi-optimization process

It is best if the ligands in datasets span a large range of physical properties such as molecular weight, number of rotatable bonds, number of hydrogen-bonding groups, etc. This allows developers to examine numerous factors that influence scoring. Furthermore, experimentally verified, inactive compounds with similar chemistry should be included, but no "assumed inactives." The original DUD set[15] is all assumed inactive compounds, and it is known that some of the original "duds" are misclassified and actually active.* In the updated DUD-E set,[16] only half of the "duds" are assumed inactives. However, it is important to note that Shoichet's work on off-target binding[17] has shown that ligands that are chemically similar to a target's native ligands have a 50% chance of binding. Through extension, could many of the assumed inactives in the literature actually be active? If so, this leads to an over-inflated enrichment problem, which can only be rectified by experimentally verified inactive compounds.

For the active compounds, their affinities should evenly span ≥4 orders of magnitude to reduce the influence of experimental error. Kramer et al[18] elegantly derived that the maximum Pearson R for a dataset[18] is dictated by the range of data and the accuracy of the experiments (see equation in Figure 2). If a model fits the data better than R_{max}, it is likely the result of over-fitting the data with too many parameters, thought it could also be sheer luck. The Pearson R from that lucky/overfit model would have a high 95% ci to indicate the low statistical significance.

Kramer et al[18] also analyzed ChEMBL[19] with a specific focus on molecules that had affinities measured in multiple, independent labs. They found that the standard deviation over that subset of ChEMBL was $\sigma_{expt} = 0.54$ pK_i. Anecdotally, my experimental collaborators consider 3-fold differences in independently measured K_i values (or K_d or

*Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: demanding evaluation kits for objective in silico screening – a versatile tool for benchmarking docking programs and scoring functions. J. Chem. Inf. Model. 2011, 51, 2650–2665

IC_{50}) to be good agreement, which translates to a very similar value of 0.48 pK_i. This makes the error from lab to lab ~0.5 pK_i, which is higher than the error bars reported in the literature (which are measures within one lab). We can rearrange the equation in Figure 2 to derive the range of data needed, based on the desired R_{max}:

$$\sigma_{data}=\frac{\sigma_{expt}}{(1-R^2_{max})^{1/2}}$$

We know that a normal distribution of data has ~95% of the data within 2 standard deviations of the mean, so we can approximate the needed range using 4 × σ_{data} and σ_{expt} = 0.5 pK_i:

$$\text{Normal Dist: Needed Range} \approx 4\times\sigma_{data}=\frac{4\times 0.5}{(1-R^2_{max})^{1/2}}=\frac{2}{(1-R^2_{max})^{1/2}}$$

Kramer et al[18] based their derivation on normally distributed experimental error and normally distributed data, but all of the rules they used for covariance and variance still hold true for uniform distributions, which means that the same equation for R_{max} can be used. It is still appropriate to use normally distributed experimental error, as that assumption has not changed. Though the standard deviation of uniform data is calculated differently, it is fortunate that it is directly proportional to the max–min range of data used. Any basic statistics text shows that the variance of a uniform distribution is $\sigma_{data}(\text{uniform})^2$ = (max–min range)^2/12. Therefore, the range we seek is directly related to σ_{data}(uniform), which is still directly related to σ_{expt} and R_{max}:

$$\text{Uniform Dist: Needed Range}=\sqrt{12}\times\sigma_{data}(\text{uniform})=\frac{\sqrt{12}\times\sigma_{expt}}{(1-R^2_{max})^{1/2}}=\frac{\sqrt{12}\times 0.5}{(1-R^2_{max})^{1/2}}=\frac{1.7}{(1-R^2_{max})^{1/2}}$$

Clearly, the range needed is slightly less if the data is evenly distributed from high to low values. If R_{max} = 0.95, then $R_{max}{}^2$ = 0.9 and the range needed is 5.4 pK_i for uniformly dist data. If R_{max} = 0.89, then $R_{max}{}^2$ = 0.8 and the range needed is 3.8 pK_i. This is why we recommend at least 4 orders of magnitude. Of course, the field typically has to use datasets with a smaller range of affinities, which would lead to much smaller R_{max}. What rescues R_{max} for smaller datasets is that they tend to be based on experimental studies from one lab that tend to have lower σ_{expt}. Extrapolating models built on that data to new data on the same system from a different lab will be limited by lab-to-lab agreement.

Lesson 5: Please stop using FXa as a model system

For a discussion of this point, please see our paper on the 2014 Exercise in this issue.[6]

Conclusion

The 2014 CSAR Benchmark Exercise is our last exercise. There was one last CSAR dataset of Hsp90 structures and affinities that was donated by colleagues at Abbott (now AbbVie) and augmented by in-house experiments at Michigan. That data has been passed on to the Drug Design Data Resource (D3R, www.drugdesigndata.org), a new effort for docking and scoring data that is headed by Rommie Amaro, Mike Gilson, and Vicki Feher at the University of California, San Diego. I recently attended their first workshop where participants discussed their outcomes for docking and scoring, using the Hsp90 set and another system from Genentech.

I was very impressed by their push for the community to develop accurate automated pipelines for docking and scoring, much like the protein-folding community has adopted. It is an important direction for our field to go if we want to make docking and scoring a robust tool that is not limited to the expert user’s intuition. Furthermore, automated pipelines would make advanced preparation of structures uncalled-for and improve reproducibility across the field. Our last exercise[6] and results from other contests[20–22] have pointed to some negative issues about setting up protein-ligand systems for participants to use. Everyone’s method is calibrated to their own setup, not someone else’s. This introduces unfair bias that would be eliminated if pipelines were accurate and improved.

Lastly, the D3R workshop was well attended and included sessions from participants of the SAMPL 5 Challenge. SAMPL 5 was based on prediction of host-guest binding free energies and estimation of water-cyclohexane partition coefficients for druglike ligands. It was a terrific idea to include a wider community interested in binding calculations and solvation phenomena. I know that National Institute of General Medical Science’s effort to create good datasets for docking and scoring is in good hands at UCSD, and I cannot wait to see the other new developments they will introduce.

Acknowledgments

We would like to thank all depositors of data and all participants in the benchmark exercises. The CSAR Center was funded by the National Institute of General Medical Sciences (U01 GM086873).

References

1. Dunbar JB, Smith RD, Yang CY, Ung PMU, 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]
2. Smith RD, Dunbar JB, Ung PMU, Esposito EX, Yang CY, 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]
3. Dunbar JB, Smith RD, Damm-Ganamet KL, Ahmed A, Esposito EX, Delproposto J, Chinnaswamy K, Kang YN, Kubish G, Gestwicki JE. CSAR Dataset Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. J Chem Inf Model. 2013; 53:1842–1852. [PubMed: 23617227]
4. 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]
5. Smith RD, Damm-Ganamet KL, Dunbar JB Jr, Ahmed A, Chinnaswamy K, Delproposto JE, Kubish GM, Tinberg CE, Khare SD, Dou J, Doyle L, Stuckey JA, Baker D, Carlson HA. CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking and Scoring/Ranking Challenge. J Chem Inf Model. 2016; doi: 10.1021/acs.jcim.5b00387

6. Carlson HA, Smith RD, Damm-Ganamet KL, Stuckey JA, Ahmed A, Convery M, Somers D, Kranz M, Elkins P, Cui G, Peishoff C, Lambert M, Dunbar JB Jr. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. J Chem Inf Model. 2016 in revisions (ci-2015-005237.R1).

7. Kleywegt GJ, Harris MR, Zou J, Taylor TC, Wählby A, Jones TA. The Uppsala Electron-Density Server. Acta Cryst. 2004; D60:2240–2249.

8. 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]

9. Li Y, Liu Z, Li J, Han L, Liu J, Zhao Z, Wang R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. J Chem Inf Model. 2014; 54:1700–1716. [PubMed: 24716849]

10. 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]

11. Chemical Computing Group Inc. MOE. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7:

12. Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, Chilov GG. CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein–Ligand Binding Affinity. J Chem Inf Model. 2011; 51:2090–2096. [PubMed: 21612285]

13. Huang SY, Zou X. Scoring and Lessons Learned with the CSAR Benchmark Using an Improved Iterative Knowledge-Based Scoring Function. J Chem Inf Model. 2011; 51:2097–2106. [PubMed: 21830787]

14. Carlson HA. Check Your Confidence: Size Really Does Matter. J Chem Inf Model. 2013; 53:1837–1841. [PubMed: 23909878]

15. Graves AP, Brenk R, Shoichet BK. Decoys for docking. J Med Chem. 2005; 48:3714–3728. [PubMed: 15916423]

16. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012; 55:6582–6594. [PubMed: 22716043]

17. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK, Urban L. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012; 486:361–367. [PubMed: 22722194]

18. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The Experimental Uncertainty of Heterogeneous Public K_i Data. J Med Chem. 2012; 55:5165–5173. [PubMed: 22643060]

19. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012; 40:D1100–1107. [PubMed: 21948594]

20. Spitzer R, Jain AN. Surflex-Dock: Docking benchmarks and real-world application. J Comput Aided Mol Des. 2012; 26:687–699. [PubMed: 22569590]

21. Corbeil CR, Williams CI, Labute P. Variability in Docking Success Rates due to Dataset Preparation. J Comput Aided Mol Des. 2012; 26:775–786. [PubMed: 22566074]

22. Repasky MP, Murphy RB, Banks JL, Greenwood JR, Tubert-Brohman I, Bhat S, Friesner RA. Docking performance of the Glide program as evaluated on the Astex and DUD datasets: a complete set of Glide SP results and selected results for a new scoring function integrating WaterMap and Glide. J Comput Aided Mol Des. 2012; 26:787–799. [PubMed: 22576241]

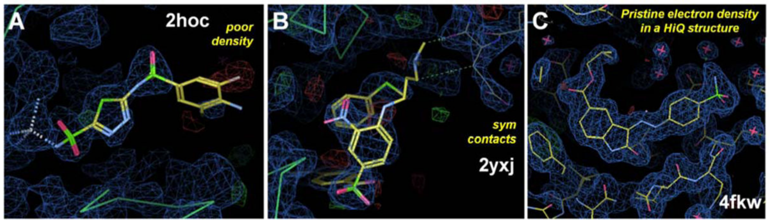

Figure 1.

Crystal structures 2hoc and 2yxj have been used as benchmarking structures in other datasets. Their structures above show missing ligand density and influential symmetry contacts. Poor density gives unjustified ligand coordinates, and crystal-packing contacts force ligands and side chains to adopt incorrect orientations. (A) In 2hoc, the ligand has very poor density for roughly half the molecule, notably the ring on the right side of the figure. (B) In 2yxj, the tertiary amine of the ligand has poor density, and it is modeled in contact with the protein in the neighboring unit cell in the upper right. (C) Structure 4fkw was produced as part of CSAR's experimental efforts, and it shows a ligand with pristine electron density for all atoms in the molecule. All side chains in the binding site are also well resolved, and even individual water molecules have good density. The ligand 62k (in the center of panel C) has an RSCC = 0.959. Ligands have yellow carbons; the protein backbone is a green chain; symmetry contacts have gray carbons. The electron density (2fo-fc map) is shown with standard blue contouring at 1.5σ. The errors (fo-fc map) are in the standard green and red colors.

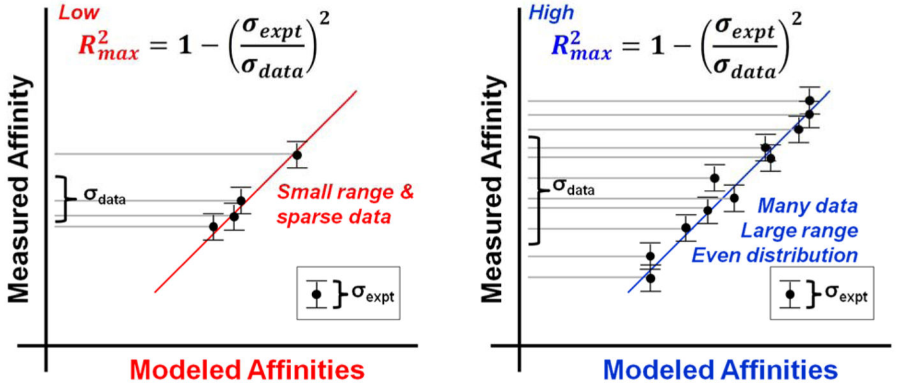

Figure 2.

The statistical limits of any model are derived from characteristics of the training data. The maximum of the model's Pearson R (R_{max}) is dictated by the standard deviation of the individual experimental values (σ_{expt}) and the range of affinities (characterized by σ_{data} over all the values in the dataset). Clearly, low error in the experiments leads to higher R_{max}. The diagram above show the effect of varying σ_{data} while keeping the same σ_{expt}. They show that weak models (red) have reduced R_{max} because of the low σ_{data} that comes from tightly clustered data with few outliers. Robust models (blue) have higher R_{max} from larger values for σ_{data} that come from adding data, extending the max–min limits of the data, and/or more evenly covering the range of data.