JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

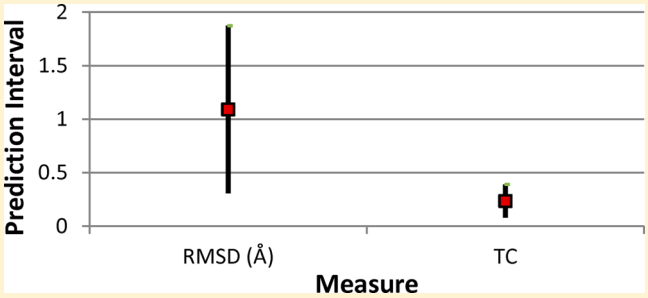
The Application of Statistical Methods to Cognate Docking: A Path Forward?

Paul C. D. Hawkins,* Brian P. Kelley, and Gregory L. Warren

OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, New Mexico 87508, United States

Supporting Information

ABSTRACT: Cognate docking has been used as a test for pose prediction quality in docking engines for decades. In this paper, we report a statistically rigorous analysis of cognate docking performance using tools in the OpenEye docking suite. We address a number of critically important aspects of the cognate docking problem that are often handled poorly: data set quality, methods of comparison of the predicted pose to the experimental pose, and data analysis. The focus of the paper lies in the third problem, extracting maximally predictive knowledge from comparison data. To this end, we present a multistage protocol for data analysis that by combining classical null-hypothesis significance testing with effect size estimation provides crucial information about quantitative differences in performance between methods as well as the probability of finding such differences in future experiments. We suggest that developers of software and users of software have different levels of interest in different parts of this protocol, with users being primarily interested in effect size estimation while developers may be most interested in statistical significance. This protocol is completely general and therefore will provide the basis for method comparisons of many different kinds.



INTRODUCTION

Cognate or self-docking is a widely reported method for assessing the quality of docking engines.[1−16] In these experiments, a ligand is extracted from a protein complex and, using a particular tool, is posed and scored back into its own binding site. The quality of the tool is then estimated by some measure of deviation between the predicted pose and the experimental pose. The underlying assumption for the relevance of such studies to the quotidian use of docking engines—where many different ligands are posed and scored in a single binding site (commonly known as cross-docking)—is that successful performance in self-docking is a necessary but not sufficient condition for successful performance in cross-docking. This is a prediction that is rarely, if ever, tested; however, testing this prediction in a rigorous way, while valuable, is beyond the scope of this paper.

We ourselves have recently reported on self-docking experiments with our docking tools.[8,17] It may then be germane to ask "why yet another self-docking study?" Our goals for this study were twofold. First, we wished to answer the question "have the changes made to our docking tools improved their performance?" with some confidence in the magnitude of the differences between older and newer versions, which required us to develop a new, more statistically rigorous method for the analysis of pose prediction data. In comparisons different types of docking methodology, the utility of self-docking experiments is concrete and immediate: if performed carefully and analyzed correctly, these experiments can provide clear-cut answers concerning the differences between methods. Second, we wished to use the pose prediction data to exemplify our general approach to method comparison, which is much more informative and rigorous than approaches currently in wide use in the community. We should note at this juncture that while we focus here exclusively on cognate docking performance, virtual screening experiments are another entirely legitimate method of evaluating docking engines. However, the levels of performance of the same tool in the two different areas can often be quite different.[6] Therefore, optimization of a scoring function for cognate docking performance will not necessarily produce corresponding improvements in virtual screening performance. The process we outline and use in this paper can be applied as-is to virtual screening data; we have used pose prediction data simply to exemplify the process and to provide concrete outputs.

Since there are different types of information to be gathered from a comparison that are of use to different interest groups, our process consists of multiple stages, each examining different aspects of the comparison. In our view, developers of tools and users of tools have different—but equally legitimate—goals when analyzing comparison data. We suggest that the minimum size of an improvement of interest to users is substantially greater than the minimum size of an improvement of interest to developers. In software development, large improvements in the function of a tool usually result from the cumulation of a number of small changes, so determining with confidence that

Received: February 20, 2014
Published: April 28, 2014

   dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

even a very small improvement has been made is important. In more technical language, *statistically* significant improvements are of interest to developers even without a *substantively* significant improvement. Still, for users, only changes that have an effect on their day-to-day work are likely to be of interest, and therefore, users will be more interested in *substantively* significant differences than *statistically* significant differences. (The existence of one of these has no implication for the existence of the other; the two are entirely independent.)

Despite the vast amount of literature published on various aspects of docking tools over the past decade, less progress has been made in this area than might have been expected. Although it is doubtless the case that substantial improvements in docking tools have been made in recent years, the data sets and methodologies used to detect and analyze these improvements have generally been insufficiently powerful. This is due, we believe, to deficiencies in three major areas of the experiments designed to assess differences between docking tools: the data set, the experimental design, and the analysis of the data generated. To address the first problem, we tested our docking tools on the carefully developed Iridium-HT (Highly Trustworthy) data set.[18] Iridium-HT is the best available data set for this kind of study for several reasons:

1. The overall structural models are easily verifiable (electron densities exist for all structures in the set).
2. The structural models are unambiguous (lacking alternate conformations for ligand or active-site residues). As such, a single ligand model best explains the experimental density, making the interpretation of deviations from that single model straightforward.
3. The ligand pose (the property we are predicting) is completely defined by the experimental electron density (vide infra for a direct application of this electron density in our method comparison).
4. There are no crystal lattice contacts with the active site or ligand. This absence of contacts with other molecules in the unit cell increases confidence that the ligand conformation is not influenced by experimental conditions (i.e., crystallization of the protein−ligand complex).
5. All of the ligand models were rerefined using the same software and protocol, providing a very high degree of consistency in the ligand models we are predicting.

We expected that a maximally reliable data set such as Iridium-HT would allow us to draw maximally reliable conclusions—more reliable than could be drawn from a set of structures simply selected from the Protein Data Bank (PDB) without due attention paid to the problems that are well-known to plague structural models in the PDB.[19]

In our experimental design, we chose to use a variety of metrics for deviation of the predicted pose from the experimental one, recognizing that no one metric could perfectly serve our purpose. We used metrics that compare the ligand model to the predicted pose directly as well as metrics that compare the experimental and predicted poses on the basis of electron density. Almost without exception in self-docking and cross-docking studies, the deviation between the predicted pose and the ligand model has been measured using the root-mean-square deviation (RMSD) based on the heavy atoms of the molecule. Since RMSD is a metric with significant problems,[20] we complemented it with a metric used in a previous study,[20] TanimotoCombo (TC), which uses global shape and chemical feature matching (see Methods for a more detailed explanation of TC).

We also eschewed comparison of the docked pose to the ligand model, as this method is predicated upon a fundamental error of type: the "experimental" pose of the ligand in the crystal structure is not experimental data but is a model developed to explain that experimental data, produced with software and guided by the human user, just like the docking pose.[21] The actual experimental data in a crystallography experiment are the structure factors from which the electron density maps are derived, and it is from fits to these maps that the three-dimensional (3D) structural model of the protein−ligand complex is constructed. Thus, a comparison of electron density synthesized from the docked pose to the experimental electron density for the ligand is more rigorous than comparing the docked pose coordinates to the ligand model coordinates because when electron densities are used, we are comparing the prediction to experimental data. Comparison of docked poses to electron density has been used previously in docking studies by Yusuf et al.[22] However, they used the comparison to electron density mostly to overcome problems with RMSD when predicting ligand model coordinates derived from partial electron density. A substantial advantage offered by the Iridium-HT data set is that every ligand is completely defined by electron density, so unlike Yusuf et al., we were able to use electron density comparison as a metric of quality for predicted poses.

Our solution to the most serious deficiency in much of the previously reported work, namely, poor data analysis, is a four-part protocol designed to extract the maximum information from comparison data and to make reliable predictions about future performance differences. In brief, the stages of this protocol are the following:

1. computation of aggregate measures of performance and their confidence intervals;
2. null-hypothesis significance testing for statistically significant differences;
3. effect size estimation for substantively significant differences;
4. quantitation of effect size and prediction interval estimation.

Comparison of many types of methods (and pose prediction methods are no exception) is often based on comparing the deviations of the predictions from the experimental data using the mean deviations (or other aggregate performance measures), with the tool exhibiting the lowest mean deviation being declared the "best".[23] This approach does have some qualitative utility, so the first stage of our process was to generate aggregate measures along with an estimation of their variability using confidence intervals. However, comparing mean or median performances directly is not very useful in quantitating the likely magnitude of the difference between methods in future experiments—which is presumably the main point of comparison studies—and is also ineffective in detecting small but real differences between tools. Therefore, in the second stage we used null-hypothesis significance testing (NHST) to determine whether a *statistically* significant difference between two methods exists (this result being mostly of relevance to the developer of the tool). If a *statistically* significant difference was found, we then, in the third stage, estimated whether a *substantively* significant difference exists (this result is of interest to the developer but is critically

important to the user who wishes to decide between the two tools). To do this, we calculated the effect size between the two methods (a unitless quantity that illustrates the magnitude of the likely difference between them; see Methods). If the effect size was of sufficient magnitude to be deemed relevant (a judgment that must be made on the basis of the individual perspective of the analyst; vide supra), we proceeded to the fourth stage of the analysis. Here we quantitated the difference between the two methods and the confidence with which this difference will be observed in the future using both an approach derived from NHST and prediction intervals for the mean difference between them. These data on likely prospective differences are obviously of great interest to the developer of the tools in question. However, quantitative estimation of the size of the difference between methods and the confidence with which such a difference might be found in the future is critical for the user to make rational decisions about the likely superiority of one method over another for a given purpose.

We used this four-stage process to compare two different posing and scoring methods within the OEDocking suite:

1. HYBRID,[8] which combines protein and ligand information in pose selection and scoring;
2. Standard docking,[17] which uses completely protein-centric scoring for pose selection and scoring.

In Standard docking, we used two different scoring functions: the Chemgauss3 function (CG3) and a recent derivative, Chemgauss4 (CG4). These two functions differ mostly in their handling of the geometry of hydrogen bonds (see Methods). Our goal was to determine with high confidence the magnitude of the difference between the performances of CG3 and CG4 and that between the performances of CG4 and HYBRID. For the first comparison, we took the perspective of the developer: do the relatively small changes in going from CG3 to CG4 make a difference in CG4's pose prediction performance? For the CG4 to HYBRID comparison, we took the perspective of the user: is the difference in pose prediction performance large enough to warrant changing from CG4 to HYBRID?

It should be noted that these two tools are entirely rigid in their operation; neither the protein structure nor the ligand conformation are changed during the posing and scoring processes. As such, HYBRID and Standard both require as input the 3D conformation(s) of the molecule to be docked. In this way, we modeled ligand flexibility in the docking process by prior computation of conformers for each molecule to be docked.

We began our investigation of pose prediction performance by redocking just the model coordinates for the ligand. While this has been done previously to probe the influence of the input conformation on the success rates of flexible docking engines,[6] here we used this approach for a different reason. Redocking of the crystallographic conformation of the ligand using rigid docking engines such as Standard removes any noise arising from the conformation generation stage of docking and focuses entirely on the ability of the posing and scoring functions to find the crystallographic pose for the crystallographic conformation of the ligand. However, this approach is reliable only when the crystallographic model is of very high quality, as it is in the Iridium-HT data set.

Having assessed the performance of our methods under these ideal conditions, we then addressed the problem of interest, cognate docking, where we investigated the ability of a given scoring function to identify a close-to-correct conformation for the ligand from the ensemble of input conformations and to place it in a close-to-correct orientation in the binding site. As mentioned above, we compared the performances of the different methods using the deviations of the docked pose from the crystallographic ligand model coordinates and from the experimental electron density to which those model ligand coordinates were fit. Unlike in most comparisons of tools for pose prediction, we carefully investigated the effect of incorporating estimates of coordinate uncertainty in the ligand model on our results and, perhaps not surprisingly, found them to make the differences between the tools compared less clear.

We were also interested in not just the relative but also the absolute performance of our pose prediction tools, but this a much more difficult problem to solve. In the evaluation of tools for virtual screening, the baseline metric of performance is itself an improvement over random selection as measured using a variety of metrics such as enrichment or area under the curve (AUC),[19] so the metric of performance is, in a sense, already absolute. However, such a clear-cut definition of random, and therefore an estimation of absolute performance, has remained obscure in the area of pose prediction. Here we propose the idea of the null model as an approach for estimating absolute performance in pose prediction. A null model is a solution to a given problem that approaches the problem in a simpler fashion than the method under study. For example, in virtual screening, a null model for similarity calculated using graph-based methods might be similarity based on atom counts,[24] while graph-based methods themselves could be considered to be the null model for 3D similarity methods. In pose prediction, we can estimate the influence of a certain term in the scoring function by comparing pose predictions from experiments using that term to predictions from experiments that do not use that term. The comparison between a method and its null model, if the null model is well-chosen, can illustrate the increase in signal the method brings over its null model. If it is found that there is no significant difference between a particular method and its less complex null model, then the more complex method should be abandoned because it provides no extra signal. Here we applied a null model to the question of the influence of hydrogen-bonding terms on pose prediction by using a purely shape-based scoring function that takes no account of chemical functionality on either the ligand or the protein. Therefore, if posing with a scoring function that includes hydrogen-bonding terms is not substantially better than that with a scoring function that does not use those terms, then the extra complexity introduced into the scoring function by the hydrogen-bonding terms is not compensated for by an increase in performance. Accordingly, the simpler scoring function should be used. An appropriate null model for the contribution of the protein structure as a whole—and not just the hydrogen-bonding interactions we consider here—to pose prediction is unclear, but we firmly believe that more effort in this area will produce substantial benefits to the field of docking as a whole.

METHODS

Software. All of the code was written in Python version 2.7, and statistical calculations were performed with SciPy version 0.1,[25] except for the power calculations, which were performed using G*Power release 3.1.7.[26] Cheminformatics functions were performed using the OEChem toolkit. Conformer generation was performed with the OMEGA toolkit, and

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

docking calculations were carried out with the OEDocking toolkit.[27]

Docking. Ligands to be docked were converted into isomeric SMILES format using the OEChem toolkit before conformer generation with the OMEGA toolkit. Following Hawkins et al.,[28] we used the following parameters for conformer generation: an RMSD cutoff (*rmsd*) of 0.5 Å, an energy window for acceptable conformers (*ewindow*) of 10 kcal/mol above the ground state, and a maximum number of conformations per molecule (*maxconfs*) of 200 for molecules with seven or fewer rotatable bonds and 800 for molecules with eight or more rotatable bonds. Docking was carried out using three different posing and scoring regimes. Two used the Standard approach, docking and scoring with either CG3 or CG4. The third used the HYBRID method, which for each ligand generates a set of poses that best match the shape and chemistry of the crystallographic ligand and then optimizes and scores those poses with the protein-centric scoring function CG4. Both CG4 and CG3 have score components that account for the shape complementarity, hydrogen bonding, ligand desolvation, and protein desolvation of a given pose. A subscore is calculated for each of these components, and the CG3 or CG4 score is a linear combination of these subscores. The only differences between CG4 and CG3 lie in the hydrogen-bonding components and the weighting of the components to produce the final score. The CG4 hydrogen-bonding component has a more detailed description of the hydrogen-bonding geometry than that of CG3. Specifically, CG4 accounts for the angle defined by the acceptor atom, the hydrogen atom, and the donor heavy atom, whereas CG3 does not. More details on these scoring functions and methods are available.[17,29]

In all cases, the docking was performed at high resolution, which, it should be noted, provides a lower bound of the resolution of a docking search of about 0.5 Å (i.e., the performance of Standard docking and HYBRID are expected to have a lower bound of around 0.5 Å on their RMSDs). We took only the top scoring pose from a method for comparison with the crystallographic structure. Receptor models for docking were generated using the make_pose_receptor (version 1.0.3) application in OEDocking version 3.0.1 for each ligand copy in the Iridium-HT data set using default settings, except that the *allowedClashes* flag was set to *allclashes*. This accounts for protein−ligand complexes with short contacts due to strong hydrogen bonds.

The Data Set. Iridium-HT has been very carefully processed to ensure that it is maximally suitable for self-docking experiments (among other tasks);[18] all of the protein−ligand complexes in Iridium-HT were rerefined and the ligands refit to their densities. Structure refinement using a single method and force field eliminates any inconsistencies in bond lengths, angles, etc., that may exist in structures drawn directly from the PDB.

The Iridium-HT data set contains 121 protein−ligand complexes and 208 separate protein−ligand instances (some ligands are found multiple times in the crystallographic unit cell). While a sample size of 208 is not the largest that has ever been used in self-docking experiments, it is larger than other well-known sets used for self-docking, such as the Astex set, which contains 85 structures.[30] An important feature of data sets to be subjected to statistical analysis is their size; all other things being equal, a larger sample size provides greater statistical power than a smaller sample size (vide infra). As mentioned in the Introduction, a substantively significant effect can be found in small data sets that are insufficiently powerful to provide a statistically significant difference. Conversely, sufficiently large data sets will provide statistical significance for differences that have no actual impact.[31] This is why our comparison protocol estimates both statistical and substantive significance.

Suitable Metrics for Comparing a Docked Pose to the Crystallographic Pose. A large number of metrics are available for comparing the pose of a ligand arising from docking calculations to the experimental pose from the protein−ligand crystal structure. The most common is a geometric measure, the root-mean-square deviation across heavy atoms (RMSD), with lower RMSD being better. While in very common use, RMSD has several well-known drawbacks.[19,20,32] One of the most obvious is its size dependence: larger molecules can exhibit larger RMSDs because of the larger number of atoms contributing to the RMSD. Another more serious problem for statistical analysis is that RMSD is bounded only on one side (its lower limit is zero, but it has no upper limit). To compensate for these and other perceived drawbacks, several other metrics have been developed.[33,34] Taking a somewhat different approach to measures like RMSD that compare poses on an atom-by-atom basis, we also deployed a global-shape-based measure, TC, which assesses the complementarity in shape and distribution of chemical features between the docked pose and the structural model in three dimensions.[35] TC, unlike RMSD, scales over a defined range (0−2, higher being better) for all molecules. While cognizant of the many drawbacks in the reliable interpretation and statistical analysis of RMSD, its geometric, atom-centered basis does provide an interpretable link between the deviation of a pose from a desired result and the coordinate error of the structure under study (vide infra).

We compared the docked result to the experimental electron density by synthesizing electron density for the docked pose using an occupancy of 1 and *B* factors of 20.0 and comparing that synthetic density to the experimental density. The match between the two was estimated using both the real-space *R* factor (RSR) and the real-space correlation coefficient (RSCC).[36] RSCCs were calculated using the rscc application distributed with AFITT version 2.3.0.[37] RSRs were calculated using unreleased code. The scales for these measures are different; RSR scales across the range 0−1 (lower being better), while RSCC scales across the range −1 to 1 (higher being better).

Handling of Model Error. All X-ray crystallography experiments have experimental error. One estimate of this error is the diffraction coordinate precision index (DPI).[38] Even though DPI is readily accessible by calculation or by download from the Uppsala Electron Density Server (EDS),[39] in only a very few studies of pose prediction has the experimental error in a structure been taken into account when analyzing the results.[40] Since the DPI is a global measure of the average coordinate uncertainty in the structure, we assumed, for lack of better information, that the coordinate error in the ligand (a very small part of the overall protein−ligand complex) is well-estimated by the DPI. We noted that the temperature factor, or *B* factor, for the atoms in the ligand is not a good estimate of the coordinate uncertainty in most of our structures. This is the case because at the resolution at which these structures were solved (95% have a resolution of >1.5 Å), the *B* factor is not an experimental observable but rather a free parameter in the refinement. We can estimate the uncertainty in the coordinates

of a structural model with the coordinate error (CE), which is derived from the DPI according to eq 1:

$$\mathrm{CE} = \sqrt{3} \times \mathrm{DPI} \quad (1)$$

When the RMSD of a predicted pose is less than or equal to the CE, there is no meaningful difference between that pose and the experimental pose. Therefore, the CE of a structure can be used to calculate a lower bound for the RMSD from docking into that structure.

In addition to using CE at the level of an individual structure, we may also use it to account for experimental error in our docking predictions at an aggregate level by using the variance-weighted mean.[20] Here the influence of an individual prediction on the final mean value is more highly weighted when the prediction is against experimental data with lower error. The equation expressing the variance-weighted mean, $\bar{m}$, is

$$\bar{m} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad (2)$$

where the weight w_i is the CE for the model with pose prediction i and y_i is the RMSD between pose prediction i and the X-ray pose.

A more detailed approach to incorporating CE into RMSD calculations at the level of an individual structure is detailed in the Supporting Information. In brief, we derived an expression for the RMSD that incorporates coordinate uncertainty in the ligand model as estimated using DPI. This new version of RMSD, which we term $\mathrm{RMSD}_{\mathrm{CE}}$, was then used to compare predicted poses to the ligand model, and the results were analyzed using the four-stage protocol detailed below.

Data Analysis. Before beginning the statistical analysis, we formed prior hypotheses about which method we expected to perform best in a pairwise comparison. To be a true prior hypothesis, this obviously must be done before the raw data are analyzed. For the comparison of CG3 and CG4, our hypothesis was that CG4 would outperform CG3, but by a small margin; the hydrogen-bonding term in CG4 would be improved over that is CG3, but all of the other components of the two functions would be the same. For the comparison of CG4 and HYBRID, we hypothesized that HYBRID would outperform CG4 by a substantial margin; HYBRID, which uses protein and ligand information together to generate and score poses, possesses a rather more complete representation of the problem than CG4, which uses only protein information.

With these prior hypotheses framed, we analyzed the data from pose prediction in four stages (shown in Figure 1). Our goal with this approach was to extract the maximum amount of information from the data, allowing us to calculate the likelihood that there is a statistically significant improvement between one method and another and to assess the likely difference in average performance between the two methods in the future. In the following, we outline the operations in and the purpose of each stage of the protocol.

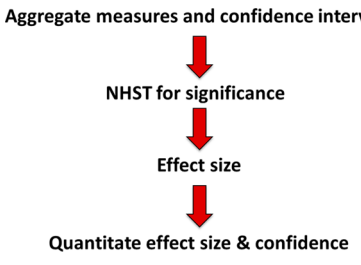


Figure 1. Protocol for method comparison.

Stage 1: Calculation of Aggregate Measures and Confidence Intervals. In the first stage of the analysis, we simply calculated aggregate measures of performance (mean and median) and 95% confidence intervals for those measures as estimates of variability. However, there are well-known general difficulties in only using summary statistics to describe data,[41] and the highly non-normal nature of RMSD data reduces the prospective utility of this analysis. Figure 2 shows

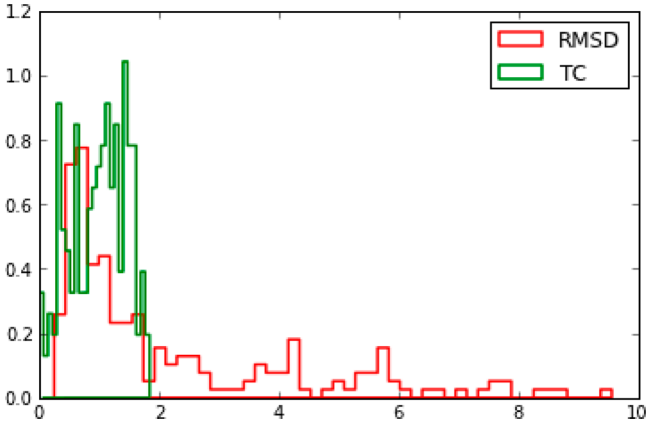


Figure 2. Distributions of RMSD (red) and TC (green) for docking with the Chemgauss4 function on Iridium-HT.

the score distributions of RMSD and TC. It can easily be seen that RMSD has a distribution that is very far from normal with a long, heavy rightward tail, while the distribution for TC is somewhat more reminiscent of a normal distribution (except that it is bounded on both sides). The RSR and RSCC distributions for docked poses, shown in Figure 3, are reminiscent of a normal distribution, except that they are both bounded above and below.

To address the problem of obtaining confidence intervals on highly non-normal data such as RMSD, we used bootstrapping.[42] According to the central limit theorem,[43] a distribution of sample means (and other aggregate measures) approaches normality as the number of samples approaches

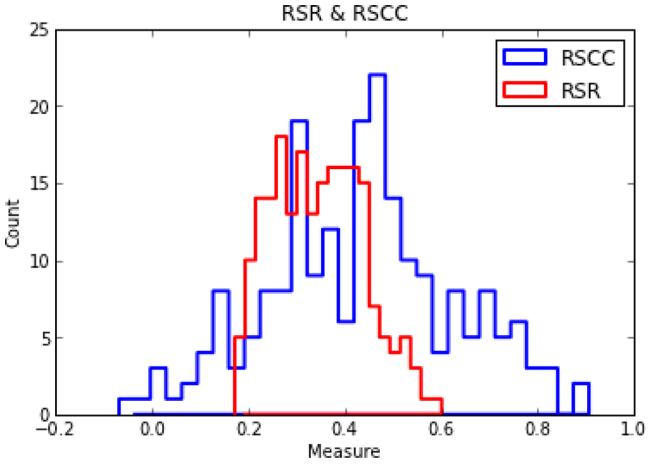


Figure 3. Distributions of RSCC (blue) and RSR (red) for docking with the Chemgauss4 function on Iridium-HT.

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

infinity. Here we obtained 95% confidence intervals for means and medians by performing nonparametric bootstrap sampling (with replacement) 20 000 times. For each of the 20 000 bootstrapped samples of the original results, we computed the mean and median. From these well-behaved (normal) distributions of 20 000 sample means or medians, we obtained their 95% confidence intervals. It should be noted that although bootstrapping, by definition, should always be performed with replacement,[44] this has not always been done. For example, in the work of Cross et al.,[6] a procedure incorrectly termed bootstrapping that omitted 20% of the data for each sample was used. This approach, while completely legitimate, is more akin to a jackknife procedure and is not actual bootstrapping.

Another issue with RMSD is that its lack of an upper bound causes interpretation problems in comparisons of both aggregate numbers and individual poses. A pose prediction of 8 Å from one docking tool is just as wrong as a prediction of 18 Å from a different docking tool, but when it is part of a larger set of predictions, the first might provide a lower mean RMSD than the second. This lower mean might lead the casual observer to conclude that the first tool with the lower mean deviation is a better pose prediction engine, which is certainly not the case.[32] Accordingly, we complemented the aggregate deviation figures with success rates by setting cutoffs for success and counting those cases where a given method passes the cutoff (lower than the cutoff for RMSD, higher than the cutoff for TC). We used two cutoffs, a stringent cutoff, C_s, and a more relaxed cutoff, C_r. The cutoffs we selected were C_s =1 Å and C_r = 2 Å for RMSD and C_s =1.5 and C_r =1.25 for TC.

This first stage provides a qualitative indication of the differences between two methods and the likely span of their future aggregate performance. To begin to quantitate the probability of a difference in performance between two methods in the future, we used the second stage of our analysis, null-hypothesis significance testing.

Stage 2: Null-Hypothesis Significance Testing. In NHST, we used the data we had gathered, D, to compare two methods using a null hypothesis, H_0. In general in NHST, H_0 is the hypothesis that there is no difference between the methods being compared or that the results from both methods are drawn from the same distribution. We rejected this null hypothesis (or asserted the alternate hypothesis, H_1, that there is a significant difference) if the test provided a sufficient confidence level. In NHST, a statistical test possesses two contrasting properties: its size or significance level, α, which is an estimation of the maximum allowable Type-I error rate (rejecting the null hypothesis when the null hypothesis is true) and its β value, an estimation of the maximum allowable Type-II error rate (not rejecting the null hypothesis when it is false). The confidence level of the test is given by $1 - \alpha$ and the power of the test by $1 - \beta$. It is common in statistical tests for the Type-I error rate to be set a priori to α = 0.05, giving a confidence level of 95%. The Type-II error rate, calculated post facto, is desired to be 0.2 or less, giving a power of 0.8 or more.[45] However, like most cutoffs, these are arbitrary and hallowed more by repetition than any unique utility. The product of a test is a p value, and H_0 is rejected if the p value is below α. Since we were interested not simply whether two methods provide different results (a two-sided test) but whether one of the methods provides superior results, we used one-sided tests in all cases. Accordingly, our null hypothesis was that one method is not better than another, and we determined the direction for the alternate hypothesis on the basis of the prior hypotheses of the performance of our docking methods stated above. Since CG4 was expected to outperform CG3, H_0 for that comparison was "*CG4 is not better than CG3.*" Similarly, because HYBRID was expected to outperform CG4, H_0 in that case was "*HYBRID is not better than CG4.*" It should be noted that as in all NHST, the p values we computed for these null hypotheses provide probabilities for finding the data (D) given the null hypothesis H_0, or $P(D|H_0)$. Unfortunately, NHST cannot compute $P(H_0|D)$, the probability that the null hypothesis is true given the data. As such, the p values we computed were not posterior probabilities for H_0 and did not address directly the likelihood that H_0 was true. This misunderstanding of NHST is regrettably common in published work.[46]

NHST on data of this kind (two methods compared on the same data set, or paired comparisons) must use tests that are robust with respect to the correlations that exist in paired data. Ignoring these correlations results in an overestimation of the variability in the data and a concomitant decrease in the statistical power of the test, $1 - \beta$, resulting in an increase in the Type-II error rate β.[47] We therefore applied tests specifically designed for the analysis of matched-pair data: the paired t test,[48] the Wilcoxon signed rank test,[49] and McNemar's test.[50] We employed both the paired t test and the Wilcoxon test because the paired t test assumes that the distribution of the differences between two methods is approximately normal while the Wilcoxon test makes no such assumption.[51] We included the Wilcoxon test here because we intended our protocol to be general and therefore made no assumptions about the distribution of the differences between the methods being compared. Since we used the paired t test and McNemar's test in somewhat unusual ways, we provide a brief introduction to these two tests.

The Paired t Test. This test operates on the differences between the predictions from two methods, which is a continuous value, and the null hypothesis for the two-sided version is that there is no difference between the methods. The paired t test assumes that the distribution of the differences between any two methods is relatively normal, and for the data reported here, that is generally true (it has been noted before[6] that differences in RMSDs are normal). The paired t statistic is calculated according to eq 3:

$$t = \frac{d - D}{\mathrm{SE}} \tag{3}$$

where d is the mean of the differences between the two methods, D is the hypothesized difference between the population means (under the null hypothesis that the methods are the same, $D = 0$), and SE is the standard error of the mean of the differences. Using the t distribution, a p value can be computed from the t statistic.

McNemar's Test. McNemar's test operates on a categorical assessment of performance, comparing those cases where one method performs well and the other badly (the discordant pairs). To define discordant pairs, a cutoff for success/failure for each metric must be specified. In pose prediction, a positive result for a method arises when its pose prediction is within an allowed deviation from the experimental pose (say, RMSD $\leq$ 2 Å) and a negative result arises when its prediction is outside the allowed deviation. There are then four possible outcomes for any prediction, which can summarized in the form of a 2 × 2 contingency table (Table 1). In this table, entries a–d are the counts of the respective outcomes; the concordant results are in

Table 1. An Illustration of McNemar's Test: The Concordant Pairs are *a* and *d*, and the Discordant Pairs are *b* and *c*

	function 1 positive	function 1 negative
function 2 positive	*a*	*b*
function 2 negative	*c*	*d*

a (both functions successfully predict the pose) and in *d* (both functions fail to predict the pose) and the discordant results (one function succeeds and one function fails) are in *b* and *c*. The null hypothesis is that p_b, the probability of finding a result in quadrant *b*, where scoring function 2 outperforms scoring function 1, is equal to p_c, the probability of finding a result in quadrant *c*, where scoring function 1 outperforms scoring function 2 (i.e., H_0 is $p_c = p_b$). McNemar's test uses the discordant pairs, *b* and *c*, to compute a test statistic and thence a *p* value for a deviation from the null hypothesis. It should be noted that the number of concordant outcomes, $a + d$, has no effect on the test result and that if $b = c$ then there is obviously no difference between the methods. Also, it should be noted that if the sum of *b* and *c* is less than 25, then the binomial variant of McNemar's test should be used. Here this was not required, as *b* and *c* always summed to more than 25. In our use of the test, we used the Yates correction for continuity.[52] McNemar's test can take one of two test statistics, χ^2 or normal (Z). The McNemar χ^2 test statistic with one degree of freedom is

$$\chi^2_{\mathrm{McN}} = \frac{(|b - c| - 0.5)^2}{(b + c)} \tag{4}$$

and the McNemar Z statistic is

$$Z_{\mathrm{McN}} = \sqrt{\frac{(|b - c| - 0.5)^2}{(b + c)}} \tag{5}$$

For the purposes of this paper, we used the McNemar Z statistic. We followed the usual procedure for this test, using a cutoff to define success that was constant across all systems studied. As is normal in docking studies, we used cutoffs of 2 Å for RMSD and 1.25 for TC based on visual inspection of poses. We also employed a cutoff for RMSD that was 2 × CE for the structure, thereby using the precision of the model structure to determine the cutoff for a successful prediction (vide infra).

We expected that TC would behave well in all of our statistical calculations. However, the lack of an upper bound on RMSD means that there could be cases in which a pose prediction was simply wrong; in these cases, direct comparisons like those done in the paired *t* test and the Wilcoxon test needed to be handled carefully. For example, if one tool produced an RMSD of 1.5 Å and the other an RMSD of 1.0 Å, then the difference of 0.5 Å in their predictions was meaningful; however, if two tools produced RMSDs of 10.5 and 10 Å, then the same 0.5 Å difference was undoubtedly not meaningful. Accordingly, we performed the paired *t* test and the Wilcoxon test on both the raw data and a subset in which at least one of the two methods under comparison produced a pose that was reasonably close to the ligand model. The criterion used to define "reasonable" was RMSD < 4.0 Å. It should be noted that since McNemar's test already uses a cutoff in defining success and failure, it is automatically robust with respect to this problem.

The results from this second stage, NHST, allowed us to determine with some confidence whether a statistically significant difference between two methods exists but not how large that difference is. The third stage of our comparison computes the effect size, or the likely magnitude of the difference between the two methods. It is only on the basis of an estimation of effect size that a decision can be made as to whether the difference is sufficiently large to warrant changing from one of the methods to another.

Stage 3: Effect Size Estimation. To estimate the effect size, or the level of difference between two methods, we used Cohen's *d* value,[53] which is computed as shown in eq 6:

$$d = \frac{(X - Y)\sqrt{(N_x + N_y - 2)}}{\sqrt{[(N_x - 1)s_x{}^2 + (N_y - 1)s_y{}^2]/2}} \tag{6}$$

where X and Y are the sample means of the results from the two methods being compared, N_x and N_y are the corresponding sample sizes, and s_x and s_y are the respective standard deviations. Cohen's *d* therefore puts the difference in the means of two sets of observations on the scale of their pooled standard deviation. Cohen placed his effect size into four classes,[53] as shown in Table 2. It should be noted that *d* is unitless and therefore can only be used to determine whether or not an effect is large enough to be worth pursuing further.

Table 2. Cohen's Classification of Effect Size, *d*

d range	$d < 0.2$	$0.2 < d < 0.5$	$0.5 < d < 0.8$	$d > 0.8$
effect size	trivial	small	medium	large

If there was at least a small effect size between two methods ($d > 0.2$), then we deemed it appropriate to proceed to the next stage, computation of the actual magnitude of the effect size and the confidence with which we would find such a difference in future experiments. If only a trivial difference was found ($d < 0.2$), then the analysis ceased; in this view, there was no utility in attempting to quantitate a trivial difference, and the two methods were therefore considered functionally equivalent, *even given that a statistically significant difference existed between them.*

Stage 4: Quantitation of Effect Size and Confidence. The fourth and final stage of our analysis aimed to convert the unitless *d* calculated in the third stage into measures useful in understanding the likely difference in performance between two methods in the future. To quantitate the magnitude of the effect size and the accompanying probability of observing data that will provide such an effect size in future experiments, we used a rearrangement of the paired *t* test from stage 2 of this protocol. In eq 3 above, a nonzero value for *D*, the hypothesized difference between the two methods, may be used to calculate the *p* value for data providing that level of difference between the two methods. Therefore, by varying *D* and recomputing the *p* value, one may determine how the *p* value varies across a range of values of *D*; plotting these data allows one to determine the probability of finding data that would give rise to any given difference *D*. This enables the user to decide on a minimum value of *D* that is of interest and then determine the *p* value for data giving that value of *D*. Also, eq 3 can be rearranged to provide an estimate of *D* at any set value of the *t* statistic. Thus, with $p = 0.05$ and $t = 1.652$ (for a large number of degrees of freedom, $n > 50$), we can compute *D* from *d* and SE, which are already known. We report this *D* (to a

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

precision appropriate to the metric in question) and the corresponding p value (which will be close to but not exactly 0.05 because of the rounding of D).

We also approached the problem of estimating future differences in performance in another way by using prediction intervals for the mean difference between two methods.[54] The 95% prediction interval for a mean difference y can be calculated according to eq 7:

$$X - t_{n-1}S\sqrt{1+\frac{1}{n}} \le y \le X + t_{n-1}S\sqrt{1+\frac{1}{n}} \quad (7)$$

where X is the mean of the differences between the methods, S is the standard deviation of the differences, n is the number of pairs compared and t_{n-1} is the value of the t statistic for $n - 1$ degrees of freedom with $\alpha = 0.025$. The 95% prediction interval calculated here provides the likely span of the difference in performance for the methods compared. If the prediction interval brackets zero, then we cannot assert with 95% confidence that the methods being compared are in fact different.

Taken together, these four stages provide all of the information required to make a rational choice between two methods attacking the same problem. The first stage, calculation of aggregate measures and confidence intervals on these aggregates, provides information on overall performance and the likely variability of future performance. The second stage, one-sided NHST, gives the probability of finding the observed data (or more extreme data) if one method is not better than the other. The third stage categorizes the magnitude of the difference between the methods on the basis of a unitless classification. The fourth stage provides an estimate of the probability of finding any given difference between two methods or the range that this difference will take at a fixed probability. The first and second stages provide information primarily of relevance to the software developer, while the third and fourth stages are required by the user to choose rationally between two competing alternatives in a fully informed manner.

Shapegauss Null Model. As a null model for the influence of hydrogen-bonding terms on pose prediction, we used the Shapegauss function (SG) from the OEDocking toolkit. The SG function poses and scores a molecule into the protein binding site purely on the basis of shape-matching with the binding site, ignoring all chemical functionality present in either.[29]

RESULTS AND DISCUSSION

We began our investigations with the simplest cognate docking experiment, redocking of the crystallographic conformation of the ligand. This experiment removed one of the main sources of noise in docking, conformer generation, and therefore allowed us to focus only on the noise generated by the posing and scoring components of the process. We performed this experiment only for the CG3 and CG4 scoring functions in Standard docking, as the HYBRID method uses the bound pose to guide scoring, which provides only trivial information on HYBRID's performance. The aggregate performance data from this experiment are shown in Figure 4. Simple inspection of the data in Figure 4 indicates that there is a very small numerical difference between CG3 and CG4 in redocking of the cognate pose of the ligand, with CG4 being slightly better (lower mean RMSD and higher mean TC), concordant with our prior hypothesis. However, when the 95% confidence intervals are taken into account, it appears that there is nothing to choose between them. We will challenge this assumption later in the analysis.

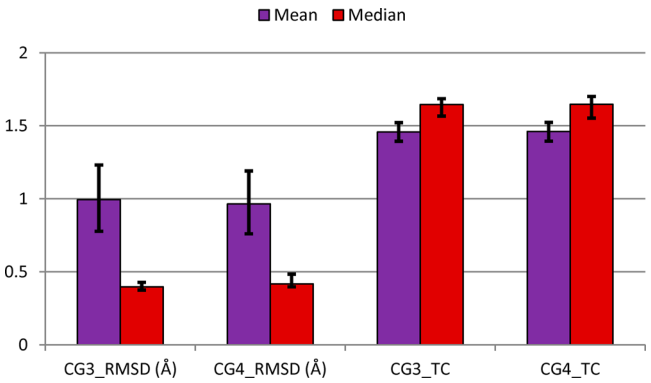


Figure 4. Mean (purple), median (red), and 95% confidence intervals for posing and scoring the crystallographic conformation of the ligand with CG3 and CG4. The deviation is measured using RMSD and TanimotoCombo (TC).

Next, we performed the self-docking experiments using CG3, CG4, and HYBRID with pregenerated conformer ensembles; this docking experiment includes noise from both conformer generation and posing and scoring. The aggregate RMSD data from the self-docking experiments are shown in Figure 5 (the aggregate data for TC are shown in the Supporting Information). We also show in Figure 6 complementary data on the success rates of the methods when cutoffs of different stringency were applied to the different measures. Two cutoffs, C_s (stringent) and C_r (relaxed), were used for each measure, as explained in Methods. Simple inspection of Figures 5 and 6 shows that, as expected, HYBRID is clearly the superior method and that CG4 appears to be slightly superior to CG3.

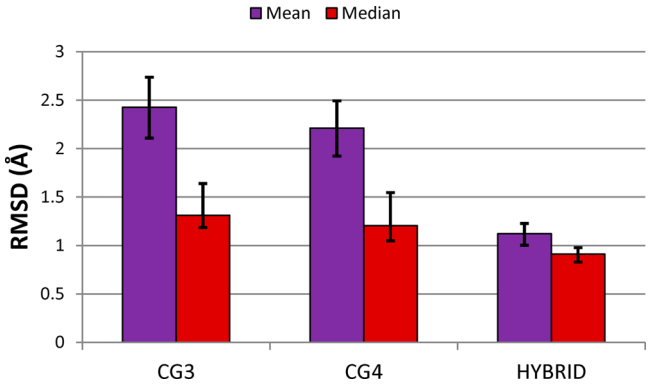


Figure 5. Mean (purple), median (red), and 95% confidence intervals for RMSD results from pose prediction with CG3, CG4, and HYBRID.

We next compared CG3 to CG4 using stages 1 and 2 of the protocol outlined in Methods and then compared CG4 to HYBRID using the same two stages.

Stages 1 and 2: CG3 versus CG4. On the basis of the numerical data illustrated in Figure 5, CG4 appears to be a better pose prediction function than CG3; it showed better (lower for RMSD and higher for TC) means and medians and smaller 95% confidence intervals for each metric, implying lower intertarget variability. Inspection of the comparisons using cutoffs in Figure 6 generally conforms to this analysis, as

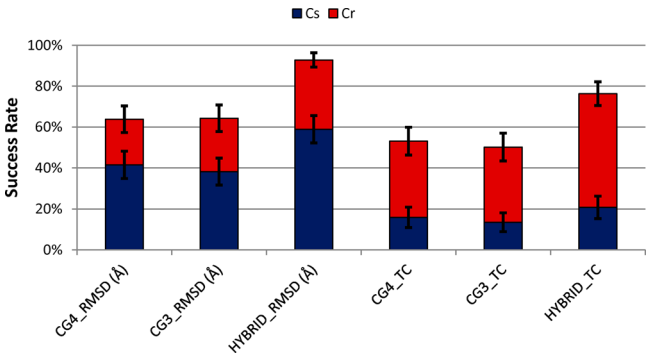


Figure 6. Percentage success rates when using cutoffs for the HYBRID, CG3, and CG4 scoring functions. C_s and C_r are stringent and relaxed cutoffs, respectively, for RMSD (1 and 2 Å, respectively) and TC (1.5 and 1.25, respectively). Error bars are 95% confidence intervals.

CG4 had the same or slightly higher success rates than CG3 for all three metrics. We have felt justified in declaring that, on average, CG4 is a better scoring function than CG3, as is often done in papers on pose prediction. This is concordant with our prior hypothesis that CG4 would outperform CG3 but by a small margin. However, in comparisons of CG3 and CG4 using either RMSD or TC, the 95% confidence intervals for any of our metrics of success overlapped substantially; therefore, using this comparison only, we could not reliably determine that CG4 is indeed a superior scoring function to CG3. To analyze the data in more depth, we applied the second stage of our protocol, NHST, to determine whether a statistically significant difference exists.

The first NHST method applied was McNemar's test, based on a categorical classification of the data (correct vs incorrect) using cutoffs for successful reproduction defined in Methods. The one-sided p values computed with the Z statistic version of McNemar's test using these cutoffs are shown in Table 3. We

Table 3. One-Sided p Values from McNemar's Test for CG3 versus CG4 Pose Predictions

	cutoff	p_Z value[a]
RMSD	<2 Å	NS
TC	>1.25	0.068

[a] p values computed using the McNemar Z statistic. NS = not significant at $p < 0.1$.

did not find a statistically significant difference ($p < 0.05$) between CG4 and CG3, which is not in accordance with our prior hypothesis. Therefore, while there is a (minor) superiority in the aggregate performance of CG4 over CG3, that superiority has no statistical significance according to McNemar's test.

We then extended the analysis using a continuous view of the data with the paired t test and the Wilcoxon signed rank test (we were justified in using the paired t test here because the distribution of differences in RMSD is quite normal; see the Supporting Information). As outlined in Methods, calculating p values using the paired t test and the Wilcoxon test on cases where both functions make poor predictions simply includes noise in the analysis. Accordingly, we performed these tests on both the complete pose prediction set, including those cases where both scoring functions produced poor predictions, and on a subset where at least one of the functions produced a pose reasonably close to the crystallographic one as defined in Methods. The results of these analyses are shown in Table 4.

Table 4. One-Sided p Values from the Paired t Test and the Wilcoxon Signed Rank Test Comparing CG3 and CG4 Using TC and RMSD[a]

	p value	
measure	paired t	Wilcoxon
all TC	**0.027**	**0.008**
TC > 0.8 (N = 92)	**0.034**	**0.004**
all RMSD	0.054	**0.033**
RMSD < 4.0 Å (N = 60)	**0.026**	**0.018**

[a] N is the number of systems surviving the cutoff. Significant results at $p < 0.05$ are shown in bold.

Here we observed a statistically significant difference ($p < 0.05$) in favor of CG4 in most cases, with TC showing significance in all comparisons and RMSD showing significance in three out of four comparisons. However, it should be noted that the cutoff of <0.05 for the p value was entirely arbitrary; it is arguable that a p value of 0.054 from the paired t test on RMSD is a notable result and should not be ignored as statistically insignificant. It was also clear that using either the complete data set or the subset from which prediction failures had been removed provided essentially the same answer. All of the results are concordant with our prior hypothesis that CG4 would outperform CG3, as they show that the data we gathered have a low probability if CG4 is not better than CG3 (our one-sided null hypothesis). The difference in the results from these continuous tests and the categorical McNemar's test is of interest, as the two tests ask rather different questions: McNemar's test examines success as defined by one arbitrary (though hopefully useful) cutoff, while the other tests examine results in a continuous manner. Thus, it could be considered that McNemar's test asks about utility (with poses passing the cutoff being assumed to be usefully close to the ligand model) while the t test and the Wilcoxon test ask about superiority. From the developer's perspective, then, the changes made in creating the CG4 scoring function have had the desired positive effect; we have high confidence that pose prediction is improved in going from CG3 to CG4, something that is not obvious simply from inspection of the aggregate performance numbers and their confidence intervals.

Having found a statistically significant difference between CG3 and CG4, a result that was not obvious on the basis of the raw data, we turned to the more clear-cut comparison of CG4 and HYBRID.

Stages 1 and 2: CG4 versus HYBRID. The data in Figures 5 and 6 show that there is a clear difference in the cognate docking performance for the HYBRID scoring function over the others: HYBRID produced better aggregate predictions and showed lower intertarget variability (smaller 95% confidence intervals). Because the HYBRID approach uses the bound ligand structure to guide pose selection, this result was entirely as expected (vide supra). Applying the second stage of our protocol, NHST, we obtained p values of 0 from all three of the tests used (Table 5). We are therefore justified in declaring that there is a highly *statistically* significant difference in performance between HYBRID and CG4. The same p values were found when subsets of the data generated using the same pose quality

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

Table 5. One-Sided p Values from the Paired t Test, the Wilcoxon Test, and McNemar's Test (Z Version) Comparing HYBRID and CG4 for Pose Prediction Using RMSD and TC

	paired t	Wilcoxon	McNemar
all RMSD	0.000	0.000	0.000
all TC	0.000	0.000	0.000

cutoffs used in the CG3 versus CG4 comparison were analyzed (data not shown).

Thus far, we have found a statistically significant effect ($p < 0.05$) showing that CG4 is a better scoring function for pose prediction than CG3 in Standard docking and a much more statistically significant difference between CG4 and HYBRID ($p = 0$). From the user's perspective, however, the following question is still open: are the improvements CG4 over CG3 and HYBRID over CG4 *pragmatically important*? The existence of *statistically* significant differences is pleasing, but whether a *substantively* significant effect exists—and with what likelihood—remains to be determined. To answer this question, we applied the third and fourth stages of our protocol.

Stage 3: CG3 versus CG4 and CG4 versus HYBRID. In the foregoing, we identified a statistically significant difference in self-docking performance between the different scoring functions and methods. As pointed out by Ziliak and McCloskey,[55] the key part of that sentence is "a difference"—we have thus far not attempted to quantitate the level of difference in performance between the functions and have thus fallen into the trap of what Ziliak and McCloskey termed "sizelessness". To express this idea another way, a difference that makes no difference is not a difference (at least in the eyes of the user of the tools), and we therefore used stage 3 of our protocol to search for substantive differences between the tools. To estimate the magnitude of the difference in performance between two methods, we used the estimate of effect size known as Cohen's d.[53] The effect sizes and 95% confidence intervals for both metrics in comparisons of CG3 versus CG4 and CG4 versus HYBRID are shown in Figure 7. On Cohen's effect size scale, the effect sizes for CG3 versus CG4 are trivial (around 0.1) while those for CG4 versus HYBRID are moderate (0.6–0.7). We thus see here a single example of the independence of statistical and substantive significance mentioned in the Introduction: the difference between CG3 and CG4 is statistically significant but not substantively

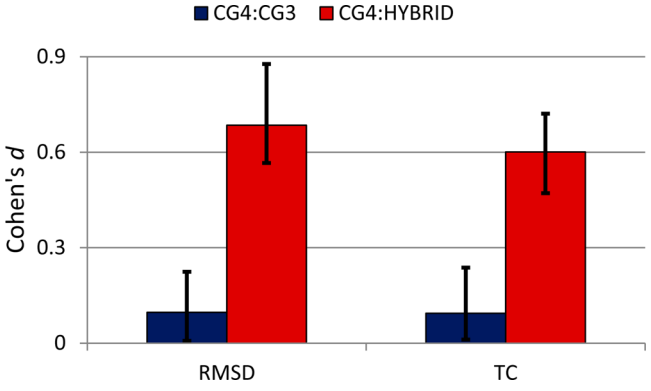


Figure 7. Cohen's d to estimate the effect size for RMSD and TC in comparisons of CG3 versus CG4 (blue) and CG4 versus HYBRID (red). The error bars represent 95% confidence intervals.

significant. Therefore, while we have discovered a statistically significant difference between CG3 and CG4, that difference, disappointingly, will likely have no effect in "real world" applications. The finding of only a statistically significant effect should not be taken to imply that the alterations made to the CG3 function to produce the CG4 function were fruitless from the standpoint of software development. Rather, the existence of that difference implies that the changes made to produce CG4 from CG3 were definitely not negative in their impact and can serve as the basis for further improvements. However, the effect size between CG3 and CG4 is so small that while numerically CG4 is a superior function to CG3 on the basis of the differences in the mean and median performance and the p value for the comparison, this superiority is pragmatically unimportant. Accordingly, we did not analyze the CG3 versus CG4 comparison data any further.

However, the difference between HYBRID and CG4 is of moderate size on Cohen's scale and is therefore likely to have an effect on the choice for use in day-to-day work. To quantitate the superiority of HYBRID over CG4, we applied the fourth stage of our protocol, quantitation of the likely difference between methods.

Stage 4: CG4 versus HYBRID. Here we compared CG4 and HYBRID using the rearrangement of the paired t test explained in Methods as well as using prediction intervals. The data generated from the paired t test (Table 6) show the

Table 6. Maximum D and p Values from the Paired t Test Estimating the Performance Difference between CG4 and HYBRID Using RMSD and TC

	RMSD	TC
D	0.85 Å	0.18
p value	0.032	0.015

highest D that gave a p value equal to or less than 0.05 and the corresponding p value for both of the metrics used to compare HYBRID and CG4. For RMSD as the measure of pose deviation, the data in Table 6 show that there is a p value of 0.032 for the data showing HYBRID outperforming CG4 by 0.85 Å on average. It should be noted that the D value is less than the difference in the means for these methods (1.1 Å; this was also true for TC) and that a p value was assigned to the data giving rise to this difference. This is a much more rigorous way to compare aggregate performance between two methods than simply comparing their mean performance. The relationship between D and the p value for these same RMSD data is plotted as a continuous function in Figure 8, allowing the user to decide upon a performance difference that is significant in his or her work and then to extract the p value for the data that would provide that difference. A similar plot can be generated for TC (not shown).

In a related approach, we generated a prediction interval for the mean difference between the two methods, as outlined in Methods. These prediction intervals are shown in Figure 9. On the basis of these data, we are 95% confident that the true mean RMSD difference between HYBRID and CG4 lies between 0.3 and 1.87 Å (one should note that it is not appropriate to say that there is a 95% probability that the true mean RMSD difference lies between these limits). The large width of the prediction interval highlights a substantial difficulty in analyzing results from docking experiments (both pose prediction and virtual screening): the high level of intersystem variability. The

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339–1355

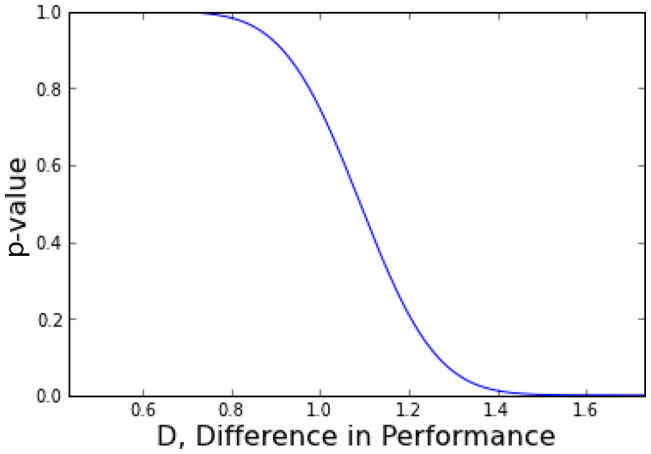


Figure 8. Plot of D vs p value from the paired t test estimating the performance difference, D, between CG4 and HYBRID using RMSD.

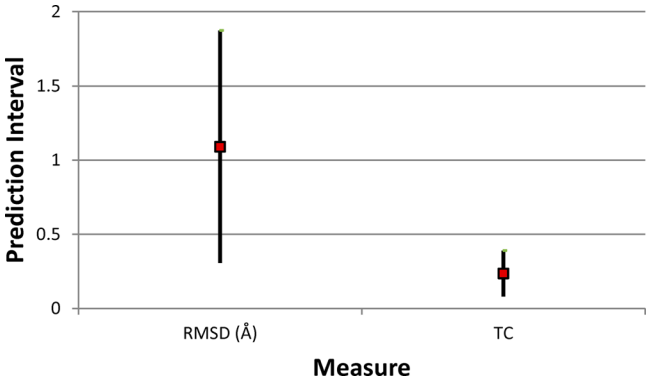


Figure 9. 95% prediction intervals for the mean differences between CG4 and HYBRID using RMSD and TC.

noise arising from this variability masks some of the inherent differences between methods. It is noteworthy that the prediction intervals for all of the CG3 versus CG4 comparisons bracket zero (data not shown), supporting the conclusion drawn from the effect size calculation that CG3 and CG4 are functionally identical in performance.

It is only on the basis of this sort of information that rational decisions about the costs and benefits of one method over another can be made.

Having completed the core analysis and gathered some firm conclusions, we then analyzed aspects of the study that could be improved.

The Problem of Power. A question we had not yet addressed in our NHST analyses was how likely our tests were to find statistically significant differences in future comparisons between methods with small differences, such as CG3 and CG4. The Iridium-HT data set, while larger than or comparable to other data sets used in self-docking experiments to date, is not enormous, leading to the possibility that Iridium-HT might not have sufficient statistical power to discriminate between possible alternatives that show small differences in performance. The fact that we did find a statistically significant effect in this comparison by no means guarantees that we would find one in a future experiment with similar data (as outlined in Methods, the rate at which we fail to find a statistically significant effect when one exists is set by the Type-II error rate, β). To estimate the Type-II error rate for the CG3 versus CG4 comparison, we estimated the power $(1 - \beta)$ of the tests we had performed comparing CG3 and CG4. The power achieved in this comparison ranged between 0.55 and 0.40 depending on the metric used, giving Type-II error rates, β, between 0.45 and 0.6 (or 45−60%). Therefore, using a data set of the size of Iridium-HT to investigate comparisons like those between CG3 and CG4, one would be expected to miss statistically significant relationships, when they exist, at a rate of 45%−60%. This is a very high Type-II error rate, recalling that the standard Type-II error rate is 20%. (In contrast, the power for the CG4 versus HYBRID comparison was always 1.00, giving a Type-II error rate of 0). We were therefore quite fortunate in our analysis to have found a statistically significant difference between CG3 and CG4. The dangers of this sort of post hoc power estimation are well-known,[56] but it does provide a useful indication of the likely reliability of the experimental design for future experiments.

We also performed a power analysis to determine the size of a data set that would be required to discriminate between CG3 and CG4 with high power given the effect sizes that we observed. To perform this, we fixed the Type-I error rate at α = 5% and set the Type-II error rate at β = 20% (the usual setting for this parameter), 10%, or 5% (giving a power of 0.8, 0.9, or 0.95). For each of our three metrics, we calculated the size of the data set required to reach that power. The results are shown in Figure 10.

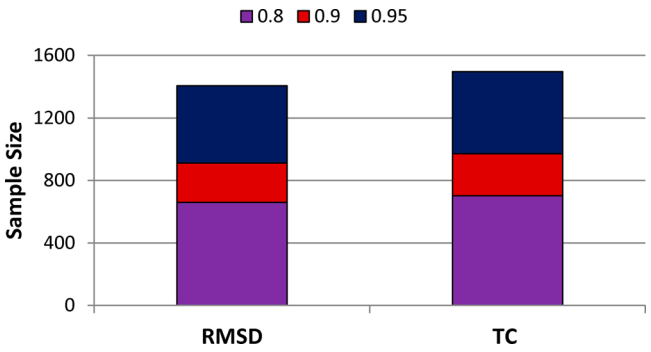


Figure 10. Sample data set sizes required to provide p = 0.05 for our metrics of success RMSD and TC at three levels of power: 0.8 (purple), 0.9 (red), and 0.95 (blue).

The sizes of the data sets required to achieve a power of 80% or better are large. For example, Figure 10 shows that a data set of 659 complexes would be required to identify that a statistically significant (p < 0.05) difference between CG3 and CG4 exists with a power of 80% (using RMSD as the metric), while a data set of 1152 complexes would be required to find a statistically significant (p < 0.05) difference between CG3 and CG4 based on RMSD with a power of 95%. Unfortunately, for the tests we employed, the Iridium-HT data set, which contains 208 separate protein−ligand instances from 121 protein−ligand complexes, is substantially underpowered for reliable comparisons between methods that are as similar as CG3 and CG4. To reiterate, what we have computed here is the reliability or reproducibility of finding a statistically significant result; power analysis assumes the existence of a statistically significant result and then computes the likelihood of finding that significant result in future experiments.

To this point in the analysis, we have omitted an important aspect of using crystallographic structure models: the comparisons performed were conducted taking no account of

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

the experimental error in the crystal structure data. It is important to determine whether this error has an effect upon our conclusions.

Incorporation of Experimental Error. In our calculations so far, we assumed (as is almost uniformly done in pose prediction) that the model ligand coordinates we attempted to reproduce were absolutely accurate and infinitely precise in their location. However, as detailed in Methods, there is a known and accessible global error in the model coordinates, the DPI, from which we can estimate the average coordinate error (CE) in the structure. The distribution of coordinate errors in the complexes in the Iridium-HT data set are shown in Figure 11.

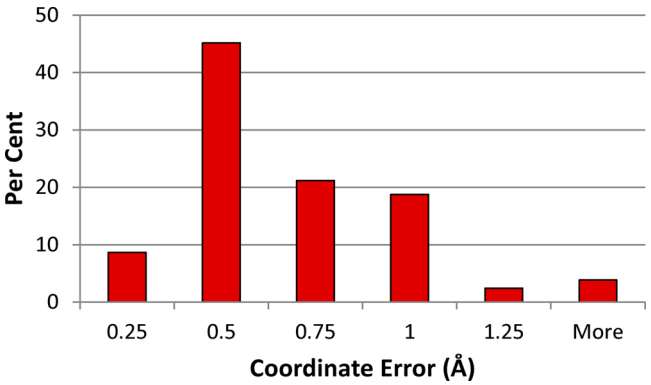


Figure 11. Distribution of coordinate error (CE) in the Iridium-HT data set.

Using this CE as an estimate of the experimental variance in the ligand coordinates, we computed both the unweighted and variance-weighted means for the RMSD of the best scoring pose for the three scoring functions along with their 95% confidence intervals (see Figure 12). In all cases, the weighted

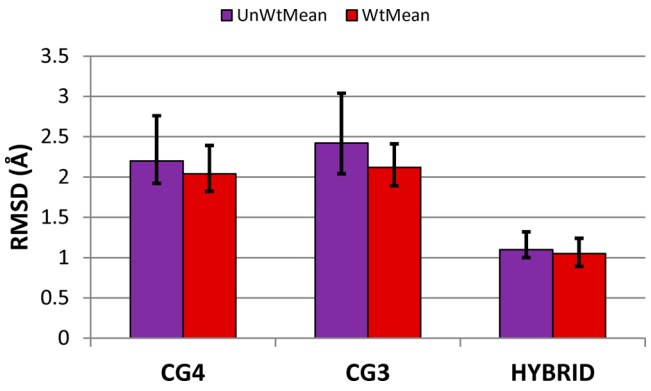


Figure 12. Weighted means (WtMean, red), unweighted means (UnWtMean, purple), and 95% confidence intervals for RMSDs for CG4, CG3, and the HYBRID method.

means were lower than the unweighted means, implying that better docking results are found for structures with lower CE. This was a satisfying outcome, as we hoped that in general computational tools would perform better with better input data, though the effect in this instance was very weak.

To take coordinate error into account at a simple level, we replaced the uniform 2 Å cutoff defining a successful pose prediction with a cutoff based on the known experimental uncertainty of the structure (2 × CE), allowing us to take appropriate account of the different levels of precision in the different models. Replacing the standard 2 Å cutoff with a cutoff of 2 × CE changes the question from "does the docking program produce a pose that is useful?" to "does the docking program produce a pose that is correct within the experimental error?" Using a cutoff based on CE applies a more stringent criterion for reproduction of the more precise crystallographic models and a less stringent one for the models with higher errors. The success rates for the different pose prediction methods with the two cutoffs are shown in Figure 13. Qualitatively, both cutoffs give the same indication: CG4 is marginally better than CG3, while HYBRID substantially outperforms both CG4 and CG3.

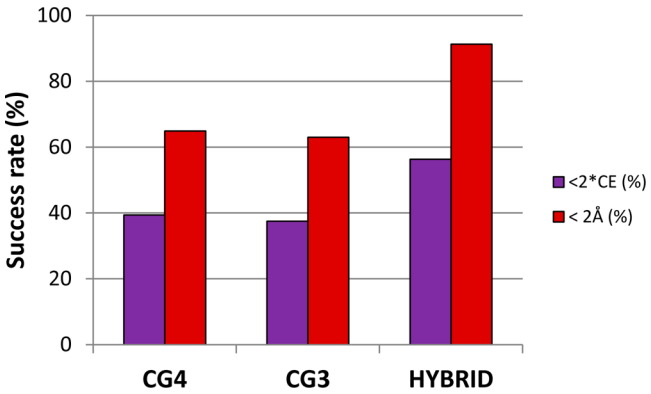


Figure 13. Percentage success rates for pose prediction for the three methods based on either a fixed 2 Å cutoff or a cutoff equal to twice the CE for the structural model into which the ligand is being docked.

In McNemar's test, we used a very similar approach: replacing the uniform cutoff of 2 Å with a cutoff of 2 × CE. In Table 7, we show the p values for the McNemar Z statistic

Table 7. Discordant Pair Analysis for CG3 versus CG4 Pose Prediction Using a Fixed Cutoff and a Cutoff Based on Coordinate Error in the Structure

	cutoff	p_Z value[a]
RMSD	2 Å	NS
RMSD	2 × CE	0.09

[a] p values computed using the Z statistic from the modified McNemar test. NS = not significant at $p < 0.1$.

using either a standard 2 Å RMS cutoff or a cutoff of 2 × CE for the structure. None of the p values were significant at the traditional $p < 0.05$ level, while at $p < 0.1$ the Z score variant showed CG4 outperforming CG3. Using this method, we obtained results for the comparison of CG4 and HYBRID that are similar to those obtained without the correction for experimental noise ($p = 0$ in all cases; data not shown).

However, the approaches above all assumed an infinitely precise location for the coordinates of the experimental pose and the predicted pose, an assumption that is clearly at odds with the existence of experimental coordinate error. As a result of this experimental uncertainty, the calculated RMSD between the crystallographic pose and a pose prediction is actually a lower bound for the "real" deviation, as the crystallographic pose is not precisely experimentally defined (see Methods for further explanation). Expanding on this concept, we investigated incorporating experimental noise into the paired t test in

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

the following way: if the difference between the RMSDs from the two methods was less than or equal to the CE of the structure, that difference was set to 0. Therefore, only those differences in pose prediction that were greater than the coordinate error of the structure into which the prediction was being made were used in the calculation of the t statistic. This approach gave a p value for CG3 versus CG4 of 0.055, practically identical to the p value computed without inclusion of experimental error (0.054).

We also compared the performance of the posing methods using $RMSD_{CE}$, a measure of RMSD that incorporates model coordinate error (see the Supporting Information for details of the calculation of $RMSD_{CE}$). This approach provided no statistically significant differences between CG3 and CG4, nor did it change the effect sizes between the methods. Therefore, perhaps unsurprisingly, trying to account for coordinate error in the ligand model caused our results to be less definite.

In the foregoing, we examined relative performance of two methods. We will now discuss a measure of absolute performance using comparison to appropriate null models.

Null Models. In considering experiments of this kind to determine the suitability or superiority of a given method, comparison to a simpler experiment is usually ignored; the question "how well would we do if we did not do what we are doing?" is rarely posed and even more rarely answered. In traditional methods of pose prediction (e.g., CG3/4), protein information is central to the calculation and ligand information is ignored (although the disposition of the protein active site is clearly biased by the presence of the bound ligand[57]). Considering the comparison between CG3 and CG4 more directly, where the differences between the two functions lie in their handling of hydrogen-bonding interactions, a null model for them could be to use a scoring function that does not use hydrogen-bonding interactions at all, such as the Shapegauss function described in Methods.

The Shapegauss Null Model. On the basis of its lack of chemical feature matching, we expected that the Shapegauss function would perform worse on average than the CG3 and CG4 functions. The pose prediction data shown in Figure 14

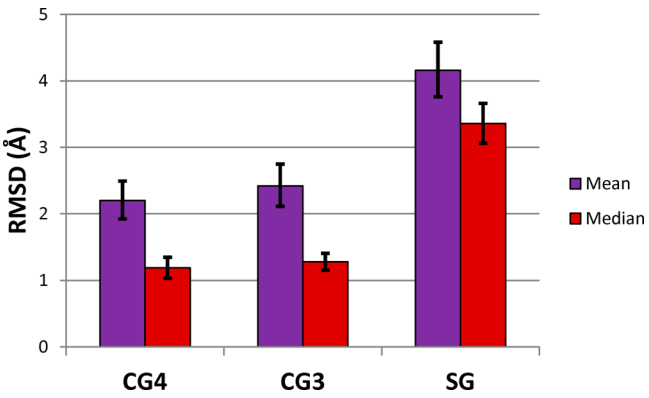


Figure 14. Means, medians, and 95% confidence intervals for RMSDs from CG3 and CG4 vs the Shapegauss null model (SG).

bear this out unambiguously. While it was scarcely necessary, we used the paired t test and the Wilcoxon test to compute p values for our data. In both cases, we found a highly statistically significant difference between the methods ($p = 0$), implying that the data were certain not to be found if SG and CG4 performed the same in pose prediction. As we did for the scoring function comparisons, we estimated the effect sizes for CG4 versus Shapegauss (see Table 8). In Cohen's scheme, the effect sizes are large for TC and medium for RMSD (though both effect sizes are close to the cutoff of 0.8 between medium and large effects).

Table 8. Cohen's d To Estimate the Effect Size for RMSD and TC when Comparing CG4 and the Shapegauss (SG) Null Model

	RMSD (Å)	TC
CG4 vs SG	0.784	0.823

Having found a substantive effect, we quantitated the effect size D and computed the p value for the data giving rise to that D using the paired t test for the three metrics (see Table 9).

Table 9. Maximum D and Derived Confidence Levels from the Paired t Test Estimating the Performance Difference between CG4 and SG Using TC and RMSD

	RMSD	TC
D	1.7 Å	0.31
p value	0.045	0.043

Again, we noted that this difference D of 1.7 Å is less than the difference in the means for CG4 and SG (1.96 Å), as was also the case for TC. We were thus confident that our more complex scoring functions possessed a significant and substantive superiority over the SG function.

Comparison Based on Electron Density. In the preceding analyses, we compared the predicted pose from docking to the model ligand coordinates using a variety of measures. As explained previously, we made a fundamental error of type in this comparison: we compared the outputs from two pieces of software—crystallographic refinement software and docking software—to each other and defined one of them (the output from the crystallographic software) to be "correct" and any difference from that "correct" answer predicted by the docking software to be undesirable. When judging the performance of any prediction method, it is usual practice to compare the predictions to experimental data. Thus, we should have compared the prediction from our docking tool to experimentally derived data, in this case the electron density that defines the ligand orientation in the binding site. We measured the similarity between the experimental electron density and the synthetic computed density for the docked pose using real-space correlation coefficient (RSCC) and real-space R factor (RSR).[36,37]

The battery of paired tests applied to the deviations from the ligand model coordinates were applied to the RSCC and RSR data, but no statistically significant differences between CG3 and CG4 could be detected (see Table 10). Again, however, HYBRID was found to be statistically significantly different from CG4 (data not shown). The lack of even a statistically

Table 10. p Values from Tests Comparing CG3 and CG4 for Pose Prediction by Comparison to Electron Density Using RSCC and RSR

	paired t	Wilcoxon
RSCC	0.696	0.643
RSR	0.248	0.361

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

significant difference between the pose predictions from CG3 and CG4 when comparing their predictions to something close to the experimental data points up a fundamental issue when using model coordinates to compare the performances of docking tools: the model coordinates are, to a large degree, too precise for the data from which they are derived. At the resolution of most structures in the Iridium-HT data set (and in the PDB as a whole), multiple slightly different solutions for the positioning of ligands that are equivalent in terms of their fit within their electron density are possible.[18,22] Therefore, comparison of a calculated pose to just one of this ensemble of possible solutions is at best systematically arbitrary and at worst close to useless.

Another contribution to the lack of differentiation between methods is fragility in the metrics themselves. This is shown in Figure 15, where the RMSD between a calculated pose and the

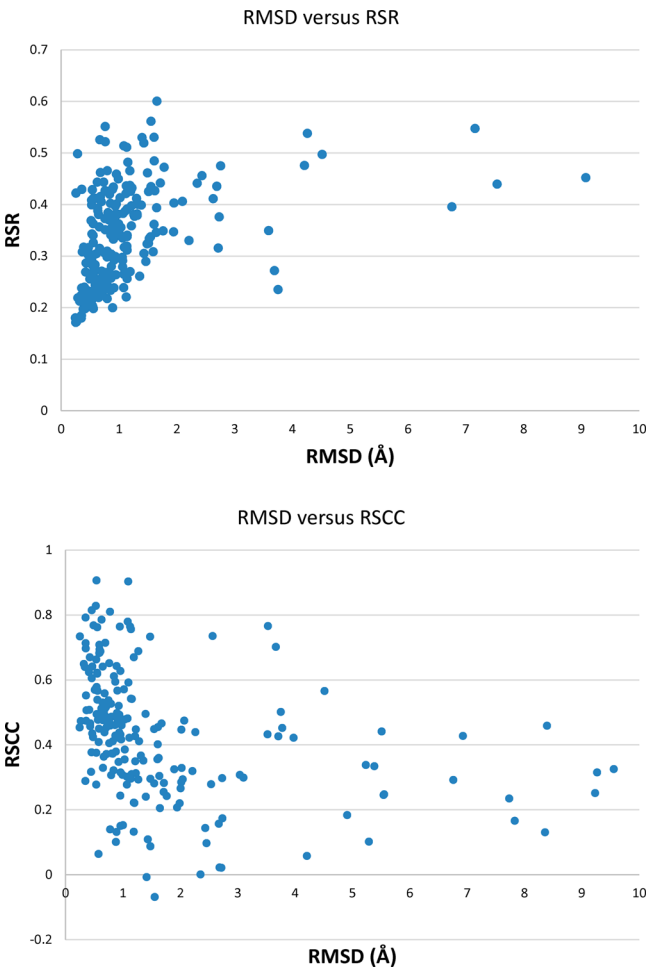


Figure 15. RMSD between a top-scoring pose and the crystallographic pose plotted against the similarity in the poses' electron densities as measured by (top) the real-space *R* factor (RSR) and (bottom) the real-space correlation coefficient (RSCC).

crystallographic model pose is plotted against the similarity in electron density between the two computed using RSR or RSCC. It is clear that RSCC and RSR are both imprecise measures of the deviation between a calculated pose and the experimental data, as they correlate very poorly with RMSD. This fragility in the metrics currently available for comparing electron densities is a problem for both crystallography and pose prediction experiments that remains unsolved.

SUMMARY

Our goal for this study was twofold: first, to compare methods for pose prediction to determine which method is superior and by how much; second, to present an approach to method comparison that could be used in a wide variety of problems, not just docking or pose prediction. We compared the HYBRID methodology, which uses both protein and ligand information in pose generation and scoring, with Standard docking, which is purely protein-centric. Within Standard docking, we used two protein-centric scoring functions, ChemGauss3 (CG3) and a (hopefully) improved version, ChemGauss4 (CG4). We wished (i) to determine whether the changes made in CG3 to produce CG4 were positive and, if so, by how much; and (ii) to estimate how much extra signal is added when pose prediction is performed using both protein and ligand information (HYBRID) over using just protein information (Standard). In order to answer these questions with maximum reliability, we set out to rectify three problems with currently accepted methods of pose prediction comparison: the data sets used, the experimental design, and the data analysis.

To address the first problem we utilized Iridium-HT, a very carefully curated protein−ligand data set with maximally reliable ligand poses. It is our view that this is the best data set currently available for this purpose, though it is by no means perfect, as we have shown here. It is worthy of note that when we repeated the experiments performed on Iridium-HT on a set of PDB model structures of lower quality (the Iridium-MT set), we found no statistically significant differences between CG3 and CG4 (see the Supporting Information). This finding further emphasizes the importance of using high-quality data sets when investigating the differences between methods that are likely to be small.

In the experimental design, we compared predicted poses to experimental ligand models by measuring the deviation between the ligand model coordinates and the predicted pose coordinates using two different measures of deviation between the ligand model and the predicted pose: RMSD and TC. We complemented this with two measures of the difference between the experimental electron density and synthetic electron density derived from the predicted pose: RSR and RSCC. In this way, we aimed to use relevant metrics of deviation that compared our predictions to both ligand models and experimental data. However, what we did not do when designing our experiments was to perform a power analysis to estimate the likely sample size we would need to reliably detect the small differences we would expect to see between two scoring functions as similar as CG3 and CG4; we found that Iridium-HT is too small a data set for this sort of purpose.

Our improved method of data analysis involved a four-stage protocol: first, comparison based on aggregate measures and their confidence intervals; second, null-hypothesis significance testing (NHST); third, effect size estimation; fourth, quantitation of the difference between the methods and prediction intervals for the mean difference. Using comparisons between the ligand model coordinates and the predicted coordinates, in the first stage we found small differences in favor of CG4 over CG3 and large differences in favor of HYBRID over CG4, entirely as we expected. In the second stage of our protocol, we found a statistically significant difference between CG3 and CG4 as well as one between CG4 and HYBRID. Unfortunately, because of the size of the Iridium-

dx.doi.org/10.1021/ci5001086 | J. Chem. Inf. Model. 2014, 54, 1339−1355

HT data set, we found that we had only about a 50−50 chance to identify this statistically significant effect, indicating that our sample size was substantially too small (the experiment was underpowered to reliably find small differences like those between CG3 and CG4).

In the third stage, we found a substantively significant difference between CG4 and HYBRID but not between CG3 and CG4. The fourth stage was applied only to the CG4/HYBRID data, as the CG3/CG4 difference was found to be too small to be worth pursuing. Thus, while we found a statistically significant improvement in going from CG3 to CG4, clearly indicating that the changes made to produce the CG4 function were positive, the magnitude of that change was so small that it would make no difference in actual use. However, for the CG4 versus HYBRID comparison, we were able to define the difference between the two methods probabilistically, allowing users to clearly understand the likely advantage that using HYBRID would provide.

Disturbingly, while CG3 and CG4 were found to be statistically significantly different when measures of deviation based on comparing ligand model coordinates to predicted coordinates were used, a comparison based on electron density found no such difference. Investigating further, we found the measures of difference in electron density to be surprisingly fragile for this purpose, showing large changes for very small changes in geometry, and therefore, while using measures that compare a prediction directly to experimental data is much more rigorous and methodologically satisfying, the measures currently available to us for comparing electron densities were found to be unsuitable. The implications of this observation remain to be fully explored.

CONCLUSIONS

Our four-stage protocol provides a framework for comparison between methods of many different kinds. We have chosen to exemplify it by application to pose prediction, but it is by no means limited in scope to pose prediction and related problems. For developers of software, all parts of the protocol can provide useful information about the usually incremental changes that arise when going from version to version. For users of software, we suggest that the most useful parts of our protocol are stages 3 and 4 related to effect size—that is, determining whether there exists a difference between two methods that would have an effect on day-to-day use. Without at least a small effect size, any statistically significant difference will remain only that—a statistical finding that will have no effect on normal practices by users (like the difference between CG3 and CG4). Therefore, we strongly advocate for estimates of statistical significance always to be complemented with estimates of effect size. As noted above, statistical significance and substantive significance are entirely independent, and from the users' perspective, substantive significance is the more useful quantity to investigate. We also advocate for estimates of the power of an experiment when considering experimental design, a step we did not undertake prior to this study, to our cost. Even though we did find statistically significant differences between the methods compared, we were just as likely to have found no effect if we had used a different data set of the same size. To avoid high levels of uncertainty in the future, we suggest that power/sample size calculations be performed as a matter of course before beginning comparison experiments of any sort in order to determine an adequate data set size based on an estimate of the likely effect size. Estimating the effect size a priori (and thence estimating the sample size), while requiring some assumptions, is vital to the appropriate design of an informative and reliable experiment.

We have remained firmly within the framework of frequentist null-hypothesis testing. This has proven productive in determining statistically and substantively significant differences between methods. However, in the NHST framework, we may only assume that the null hypothesis that we test is either true or false; the methodology does not permit the estimation of posterior probabilities for the null hypothesis, which is arguably the most useful quantity to pursue. It is likely that Bayesian methods,[58] which do estimate these posterior probabilities, can be fruitfully applied to this problem in the future.

ASSOCIATED CONTENT

Supporting Information

Python scripts to generate conformers, to perform the docking, and to calculate the statistical measures used; frequency distribution of the differences in RMSD between poses from Standard docking using CG3 and CG4; means, medians, and 95% confidence intervals for pose reproduction using CG3, CG4, and HYBRID; and procedure for calculating RMSD accounting for model coordinate error. This material is available free of charge via the Internet at http://pubs.acs.org. The Iridium data set is available for download from http://www.eyesopen.com/iridium.

AUTHOR INFORMATION

Corresponding Author

*E-mail: phawkins@eyesopen.com. Tel. 505-473-7385 ext. 65.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Matthew Geballe, Dr. Christopher I. Bayly, and Dr. Thomas A. Darden for useful discussions and insights on the practice and theory of statistics. Ms. Annie Lux is thanked for proofreading the manuscript. We thank the anonymous reviewers for their valuable suggestions.

REFERENCES

(1) Beuming, T.; Sherman, W. Current assessment of docking into GPCR crystal structures and homology models: Successes, challenges, and guidelines. *J. Chem. Inf. Model.* **2012**, *52*, 3263−77.

(2) Brozell, S. R.; Mukherjee, S.; Balius, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 749−73.

(3) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−15.

(4) Corbeil, C. R.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J. Chem. Inf. Model.* **2009**, *49*, 997−1009.

(5) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−86.

(6) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−74.

(7) Liebeschuetz, J. W.; Cole, J. C.; Korb, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 737−48.
(8) McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 897−906.
(9) Mukherjee, S.; Balius, T. E.; Rizzo, R. C. Docking validation resources: Protein family and ligand flexibility experiments. *J. Chem. Inf. Model.* **2010**, *50*, 1986−2000.
(10) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: The benchmarking results and strategies for improvement. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 675−86.
(11) Novikov, F. N.; Stroylov, V. S.; Zeifman, A. A.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 725−35.
(12) Repasky, M. P.; Murphy, R. B.; Banks, J. L.; Greenwood, J. R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R. A. 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*, 787−99.
(13) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claussen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 701−23.
(14) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 687−99.
(15) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54*, 5422−31.
(16) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative performance of several flexible docking programs and scoring functions: Enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599−608.
(17) McGann, M. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578−96.
(18) 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*, 1270−81.
(19) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: Pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−90.
(20) Hawkins, P. C. D.; Nicholls, A. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−36.
(21) Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallogr., Sect. D.: Biol. Crystallogr.* **2000**, *56*, 249−265.
(22) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411−22.
(23) Thomsen, R.; Christensen, M. H. MolDock: A new technique for high-accuracy molecular docking. *J. Med. Chem.* **2006**, *49*, 3315−21.
(24) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−75.
(25) Jones, E.; Oliphant, T. *SciPy*, version 0.1, 2013; available at www.scipy.org.
(26) Faul, F.; Erdfelder, E.; Lang, A. G.; Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175−91.
(27) OpenEye Toolkits. http://www.eyesopen.com/toolkits.
(28) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. 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*, 572−84.
(29) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.
(30) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−41.
(31) Weber, G. W.; Prossinger, H.; Seidler, H. Height depends on month of birth. *Nature* **1998**, *391*, 754−55.
(32) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein−ligand docking programs is difficult. *Proteins* **2005**, *60*, 325−32.
(33) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: A generally applicable replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49*, 1889−900.
(34) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−81.
(35) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(36) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47* (Part 2), 110−9.
(37) AFITT. http://www.eyesopen.com/afitt.
(38) Blow, D. M. Rearrangement of Cruickshank's formulae for the diffraction-component precision index. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 792−7.
(39) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240−9.
(40) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophore-based protein−ligand docking. *J. Med. Chem.* **2004**, *47*, 6804−11.
(41) Anscombe, F. J. Graphs in statistical analysis. *Am. Stat.* **1973**, *27*, 17−21.
(42) Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, 1−26.
(43) Rice, J. A. *Mathematical Statistics and Data Analysis*, 3rd ed.; Duxbury: Belmont, CA, 1995; Chapter 5.
(44) Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171−85.
(45) Mazen, A. M. M.; Hemmasi, M.; Lewis, M. F. In search of power: A statistical power analysis of contemporary research in strategic management. *Acad. Manage. Proc.* **1985**, *1985*, 30−4.
(46) Gigerenzer, G. We need statistical thinking, not statistical rituals. *Behav. Brain Sci.* **1998**, *21*, 199−200.
(47) Sainani, K. The importance of accounting for correlated observations. *PM&R* **2010**, *2*, 858−61.
(48) Student. The probable error of a mean. *Biometrika* **1908**, *6*, 1−25.
(49) Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80−3.
(50) McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153−7.
(51) http://en.wikipedia.org/wiki/Paired_difference_test.
(52) Yates, F. Contingency table involving small numbers and the χ^2 test. *Suppl. J. R. Stat. Soc.* **1934**, *1*, 217−235.
(53) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, 1988.
(54) Casella, G.; Berger, R. L. *Statistical Inference*, 2nd ed.; Duxbury: Pacific Grove, CA, 2002.
(55) Ziliak, S. T.; Mccloskey, D. N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*; University of Michigan Press: Ann Arbor, MI, 2008.

(56) Hoenig, J. M.; Heisey, D. M. The abuse of power. *Am. Stat.* **2001**, *55*, 19−24.
(57) Tuccinardi, T.; Botta, M.; Giordano, A.; Martinelli, A. Protein kinases: Docking and homology modeling reliability. *J. Chem. Inf. Model.* **2010**, *50*, 1432−41.
(58) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*; CRC Press: Boca Raton, 2003.