JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Numerical Errors and Chaotic Behavior in Docking Simulations

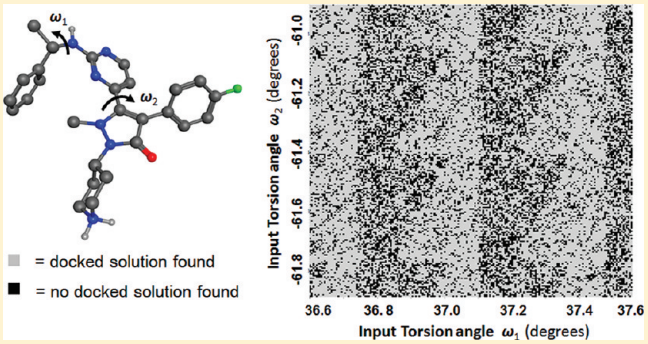
Miklos Feher*,† and Christopher I. Williams‡

†Campbell Family Institute for Breast Cancer Research, University Health Network, Toronto Medical Discovery Tower, 101 College Street, Suite 5-361, Toronto, ON, M5G 1L7, Canada

‡Chemical Computing Group, Suite 910, 1010 Sherbrooke Street W., Montreal, QC, H3A 2R7, Canada

Supporting Information

ABSTRACT: This work examines the sensitivity of docking programs to tiny changes in ligand input files. The results show that nearly identical ligand input structures can produce dramatically different top-scoring docked poses. Even changing the atom order in a ligand input file can produce significantly different poses and scores. In well-behaved cases the docking variations are small and follow a normal distribution around a central pose and score, but in many cases the variations are large and reflect wildly different top scores and binding modes. The docking variations are characterized by statistical methods, and the sensitivity of high-throughput and more precise docking methods are compared. The results demonstrate that part of docking variation is due to numerical sensitivity and potentially chaotic effects in current docking algorithms and not solely due to incomplete ligand conformation and pose searching. These results have major implications for the way docking is currently used for pose prediction, ranking, and virtual screening.



INTRODUCTION

We all learn about experimental error early on in our scientific education, but error is typically *not* considered in computational experiments, because it is usually assumed that computer calculations are completely reproducible—given the same input, the computer produces the same output, with no "error bars" on the computed result. However, reproducibility in digital calculations breaks down in many instances. The options used to compile a program, using multiple CPUs in a calculation,[1] small changes to input data, and even the order of mathematical operations in a processor can all give rise to small variations in computed output. These variations start off small but can be magnified by long iterative calculations such as molecular mechanics minimizations,[2] molecular dynamics,[3] and protein–ligand docking.[4] This magnifying of variations can cause repeated runs of identical or near-identical input to diverge and produce significantly different final results. In this work we single out docking programs to demonstrate how seemingly negligible variations in ligand input coordinates, and even changing the atom order in a ligand input file, can produce unexpectedly large variations in the docking output. Statistical methods for characterizing docking variations are described, and the implications of these results on the practice of docking are discussed.

Protein–ligand docking has become an indispensable tool in computer-aided drug discovery. Starting from a suitably prepared all-atom protein structure and a single representative ligand conformation, docking programs typically search ligand conformations and binding pocket placements to produce a list of docked poses and corresponding scores, with a single top-scoring pose as the best solution. On the basis of their underlying algorithms, docking methods can be broadly classified as either *stochastic* or *deterministic*. Stochastic docking programs employ random elements such as genetic algorithms in their searches, so repeated runs of the program with the same input will produce different lists of docked solutions with a range of top-scoring poses. This is demonstrated schematically in Figure 1a, where the input ligand structure is represented as a single point in ligand conformation space and the range of top-scoring poses resulting from repeated docking runs is represented by contour lines in docked solution space. Thus with stochastic docking programs, it already seems natural to estimate the error in the top scoring pose. Indeed, variations in stochastic docking output had been investigated,[4,5] but in general, detailed error analysis of this sort is rarely performed in docking studies.

In contrast to stochastic docking programs, deterministic programs contain no random elements, so repeated runs of the exact same input produces the exact same list of docked poses and the exact same top-scoring pose every time, as represented by a single point in the docked solution space of Figure 1b. However, we have recently shown[4] that deterministic docking programs can produce very different top-scoring poses when presented with different but entirely valid ligand input structures. Although part of docking variation arises from

Received: December 13, 2011
Published: March 1, 2012

   dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724–738

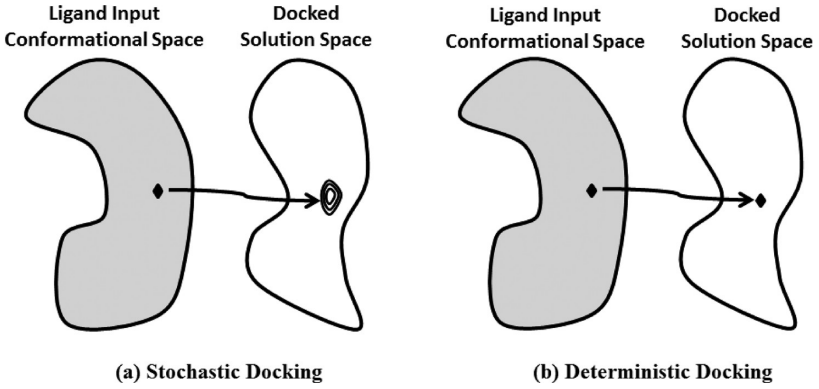


Figure 1. Stochastic and deterministic docking: expected docking results for multiple runs of identical input using stochastic and deterministic docking programs. (a) Due to randomized elements in the stochastic docking algorithms, repeated runs with an identical ligand input will produce top-scoring poses with small variations in the top score and pose, as represented by the small contoured region in docked solution space. (b) With deterministic docking programs, repeated runs of an identical ligand input structure (represented as a point in ligand input conformation space) will in every case produce an identical top score and pose, represented as a point in docked solution space.

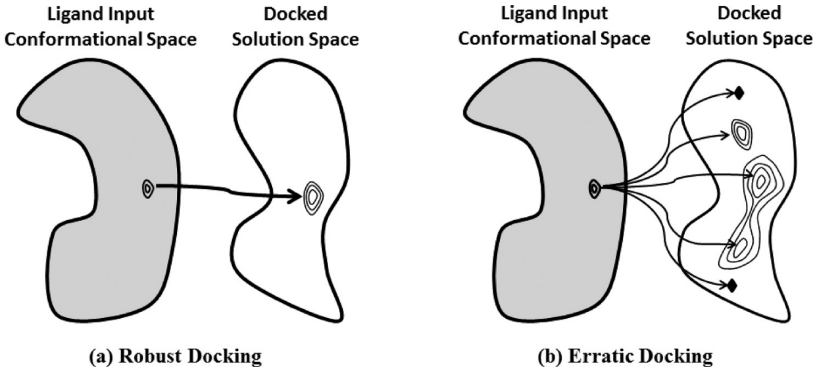


Figure 2. Robust vs erratic docking: comparison of docking behavior with respect to small perturbations in the ligand input conformation. The contour lines in ligand input conformation space represent the set of near-identical ligand input conformations used in the different runs. The contour lines in docked solution space represent the spread of top-scoring poses produced by these input conformations. (a) With robust behavior, a small spread in ligand input conformations produces a small spread in the top-scoring poses. (b) With erratic behavior, a small spread of ligand input conformations can produce wildly varying top-scoring poses.

insufficient ligand conformational searching,[6] examples where near-identical ligand input structures gave rise to significantly different top-scoring poses could only be explained as numerical sensitivities in the docking programs.

The current study seeks to better characterize docking variations that arise from tiny changes to the ligand input structure. The experimental scenario is illustrated schematically in Figure 2; each member of an ensemble of near-identical ligand input structures (represented by contoured regions in ligand input conformation space) is docked, and the top-scoring pose from each input structure is kept (represented by contoured regions in docked solution space). The diagrams in Figure 2 represent two limiting cases of *robust* and *erratic* docking. In robust docking, essentially the same top pose and score is produced from each run, so output variations are small, as represented by the small contoured region in the docked solution space of Figure 2a. In erratic docking, small changes to ligand input lead to large output variations, represented by large contoured regions and outliers in the docked solution space of Figure 2b. The variations in pose and score can be analyzed with statistical methods to characterize the degree of robust and erratic behavior and to give insight into the expected error bars associated with a docked pose and score.

In this study, the ligand input ensembles consisted of 500 conformations with identical bond lengths and bond angles, which differ only in their nonring torsion angles by a maximum deviation of ±0.1°. These conformations have average pairwise RMSDs of 0.01 Å or less and are virtually indistinguishable when displayed on a computer monitor at normal resolutions and screen sizes. These conformations are also indistinguishable from the point of molecular geometry optimizations when usual gradient cut-offs are applied. Each member of these near-identical input ensemble was docked to its target using the commercially available GOLD[7] and Glide[8] programs as examples of stochastic (GOLD) and deterministic (Glide) docking protocols. The experiments were run with three different precision settings for each software package in order to examine the effect of these settings on docking output variations. The distributions of top-scoring poses were analyzed with quartile−quartile (Q−Q) plots to assess to what degree the results follow a normal distribution, and box-plots were used to describe the score and pose variations across docking methods and targets. Normally it would be assumed that such small differences in ligand input files would have a negligible effect on the docking results, but this study will show otherwise.

METHODS

Protein Preparation. For comparative purposes, we used the same set of 10 kinase and 3 nuclear receptor targets as in our previous work.[6] The pdb codes and the names of the

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

considered targets were the following: 1ke5 (cyclin dependent kinase, cdk2), 1of1 (thymidine kinase), 1opl (c-abl tyrosine kinase), 1p62 (deoxycytidine kinase, dck), 1unl (cyclin dependent kinase, ckd5), 1t46 (c-kit tyrosine kinase), 1ywr (mitogen-activated protein kinase, p38), 1y6b (vascular endothelial growth factor receptor, vegfr2), 2br1 (checkpoint kinase, chk1), 1pmn (c-jun terminal kinase, jnk3), 1m2z (glucocorticoid receptor), 1z95 (androgen receptor), and 1sj0 (estrogen receptor-α). These targets were selected in part because the corresponding X-ray structures were carefully checked for errors.[9] These protein structures had been prepared for GOLD docking[9] and were used directly. For the preparation of these structures for Glide docking, the protein preparation wizard was applied (hydrogen bond optimization and restrained structure minimization with the OPLS-AA force field to a maximum rmsd of 0.3 Å). Next a receptor-grid was generated using default parameters (inner box with 10 Å sides that should contain ligand midpoints during docking and an outer box with 20 Å sides from the ligand centroid that should contain docked ligands, no scaling applied).

Ligand Ensemble Preparation. The kinase and nuclear receptor ligands were identical to those used in our previous works; they were the cognate ligands of the targets listed in the previous section.[4,6] Ligand preparation was also kept consistent with our previous work. Ligands were first read into MOE, protonated/deprotonated using the Wash process, rebuilt into 3D using Corina,[10] and then minimized in MOE with the MMFF94x[11−13] force field to a gradient of 0.0001 kcal/mol Å^2. Using this starting point, conformations with uniformly spaced torsional angles were generated. The same number of steps was used along each torsion angle up to a maximum of 0.1° per angle, such that the total number of conformations was greater than 500 and as close to this number as possible. This was followed by a diverse subset selection to identify the 500 most diverse conformations from this set. This ensemble represents very small torsional variations around the starting conformation with RMSDs less than 0.01 Å; the corresponding structures are so similar that when displayed together, they cannot be distinguished on a regular computer screen.

Docking. The docking methods used in this study were similar to those described previously.[6] Since our aim was to establish error distributions, the settings applied in the Glide[8] and GOLD[7] programs were generally not customized, unless otherwise stated in the text. (For Glide, these settings included sampling nitrogen inversions and ring conformations within a 2.5 kcal/mol window and minimization following placement using the Amber forcefield up to a maximum of 100 steps. In the case of GOLD, the binding site included atoms within 6 Å of ligand atoms, nitrogen inversions and carboxylic acid flips were sampled and scoring was performed with the GoldScore function allowing early termination if the top 5 solutions were within 1.5 Å from each other.) Three different Glide approaches were tested; extra precision (XP), standard precision (SP), and high throughput (HTVS). The HTVS and SP methods use the same scoring function, but HTVS is faster because it considers a reduced number of conformations and is less thorough in the final torsional refinement and sampling. The XP method uses more extensive sampling and a more sophisticated scoring function with stricter requirements for ligand−receptor shape complementarity.

Three different search GOLD efficiency settings were tested; very flexible (200%), default (100%), and library screening (30%). These settings influence the number of GA operations, which is approximately proportional to the percentages given in parentheses. To reduce the effect of stochastic variations with the genetic algorithm, GOLD runs are typically repeated 10 times by default. Experiments were also performed using 80 GA runs with both the default and the very flexible search efficiencies. It was found that increasing search efficiency and the number of GA operations both provide a small and nearly indistinguishable improvement. Thus, for simplicity, some of the results with the very flexible option are only provided with 80 GA runs. Glide and the "very flexible" GOLD calculations were run on different SunFire and HP systems with 32-bit and 64-bit architecture under RedHat Enterprise 4. The rest of the GOLD calculations were obtained on a Dell Precision T7500 running under Windows XP (64 bit).

Data Analysis. In all cases the top-scoring pose from each docking run was taken as the result. Thus for each input ensemble, 500 poses with corresponding scores were produced. The rmsd of each pose was computed as the rmsd to the top-scoring pose of the *first* conformation, i.e., as produced by docking the original unperturbed Corina input structure. The degree of normal behavior of both the score and rmsd distributions was analyzed using quartile−quartile (Q−Q) plots. The Q−Q plots correlate the actual rmsd and score with theoretical values computed from a Gaussian distribution around the mean and standard deviation of the actual rmsd and score data. (This latter was obtained for the sorted rmsd distances and scores by using the normal inverse cumulative distribution function NORMINV in Microsoft Excel that has three parameters, the mean, the standard deviation, and the cumulative probability, derived from the rank proportions of these rmsd distances and scores.) If the actual rmsd and score distributions are indeed perfectly Gaussian, they will correlate with the theoretical values, yielding a Q−Q plot which is a straight line with a slope of 1, an intercept of 0, and a correlation coefficient (R^2) of 1 between the actual values and the values computed from the normal distribution. If the actual distributions are not Gaussian, the Q−Q plots are typically nonlinear with a low R^2 correlation coefficient.

The score and rmsd distributions were also described using box-plots, which display the sample minimum (MIN), lower quartile (Q1), median (Q2), upper quartile (Q3) and the sample maximum (MAX). In these box-plots, the solutions that appear in the gray boxes span the first to the third quartiles (Q1-Q3), while the bottoms and tops of the vertical lines indicate the MAX and MIN values and display the entire range of data. Distribution of conformations were analyzed using metric scaling.[14,15] In metric scaling, the distance matrix was projected to 2D space in such a way that each conformation is represented by a point and the geometric distance between the projected points remains proportional to the original distances. The plots display the x, y coordinates of the projection that retains the relative distances between the docked poses.

(ω_1, ω_2) Torsion Plots. Input ensembles of 40 000 ligand structures were created by varying only two ligand torsion angles (ω_1, ω_2) simultaneously over a small angle range (±1°). The members of the ensemble were docked, and the top-pose was kept in each case. Docking results such as score, pose cluster, and pass/fail state were plotted as functions of the input dihedral angles (ω_1, ω_2). Since the Cartesian coordinate changes resulting from these torsion perturbations were often less than 10^{-5} Å, these perturbed input structures could not be saved to low-precision file formats such as MDL SD or PDB, because the coordinate differences would be lost. Instead, these

fine-grained input structures had to be saved to the Maestro format, which records 16 digits of precision in the Cartesian coordinates and can thus capture the small structural differences. This experiment was performed on the 1ywr and 2br1 systems using GlideHTVS. The 1yrw runs were classified as either "pass" (a docked pose produced) or "fail" (no poses produced). With 2br1_chk1, all 40 000 input structures produced a docked solution, so the output was analyzed in depth by clustering the solutions based on pose and score similarity. This produced six pose clusters, with average pairwise RMSDs less than 1.2 Å within a cluster and greater than 2.5 Å between clusters. The top-scoring poses were also assigned to bins based on their scores.

Reordering Atoms in Input Files (Atom-Permuted Files). Another way of introducing differences into ligand input files *without* changing the atom coordinates is to permute the order of the atoms in the input file. Every other attribute of the ligand input file—bonding, atom types, Cartesian coordinates—remains identical. The only difference is the order in which the atoms are listed in the input file. Atom-permuted input files were created by selecting a docking input file and creating 500 copies by reordering the atoms using the "mol_aPermute" SVL command in the MOE software. The atom-permuted files were saved to SD format for input into the docking program.

RESULTS AND DISCUSSION

As described in the Methods section, 500 near-identical copies of each ligand input structure were generated by applying small torsion perturbations ($\leq 0.1°$) to the minimized Corina structure. The 500 structures were then docked to their corresponding targets and only the top-scoring pose retained from each run. We have previously investigated the reproducibility of docking runs using *identical* input (see Table 2 of ref 4) and found that deterministic docking calculations are fully reproducible, whereas the variability for stochastic methods typically stay within acceptable limits (pairwise rmsd < 0.5 Å, score variability < 3%). Since the torsion angle differences between the input structures seem negligible in this work, one would *not* expect to see erratic docking behavior, with large variations in the top score and/or pose. However, even with these tiny ligand input differences, only a few combinations of target and docking program behaved in a robust fashion. Most of the examples seemed to end up between the robust and erratic limiting cases, with many exhibiting strikingly erratic behavior.

It is important to note here that robustness does not equal docking accuracy. Results from robust docking might be completely reproducible when using different input sources but could still present an incorrect outcome (incorrect pose or poor quality ranking). On the other hand, obtaining a "correct" result using one input source but a different result from another input is also an undesirable outcome. Clearly, calculations need to be both accurate and robust, but these are two separate properties. A detailed discussion of docking accuracy is beyond the scope of this work, but note that we have looked at accuracy for the same targets in Table 6 of our previous work.[4] We used a wider variety of input sources and our conclusion was that given the lack of robustness, there seems to be no ligand starting geometry that guarantees that the best docking pose will be produced and docking the cognate ligand in its experimental conformation as input can produce worse results than docking some random conformation.

Examples of Robust, Substantially Robust, and Erratic Docking Results. The GOLD docking to cdk2 kinase (PDB code: 1ke5), GlideXP docking to p38 (PDB code: 1ywr), and GlideHTVS docking to chk1 kinase (PDB code: 2br1) runs were chosen as examples of robust, substantially robust, and erratic docking behavior. Robust docking results, demonstrated by GOLD docking to cdk2 (PDB code: 1ke5) in Figure 3, are characterized by a relatively narrow score range (less than 10%), a single binding mode, small rmsd and score differences between the obtained solutions, and a metric scaling plot which indicates that all top-scoring poses fall within the same binding mode. The histograms in Figure 3 show the actual rmsd and score distributions relative to the *first* solution, i.e., the one obtained from the unperturbed minimized Corina structure. The Q−Q plots in Figure 3 show that score and pose rmsd distributions correlate well with the ideal values, indicating that they both follow a normal distribution. This suggests that robust docking results can, to a reasonable approximation, be described as a central pose and score with associated standard deviations. For example, the GOLD docking to ckd2 (PDB code: 1ke5) could be given a score of 58.85 ± 0.14 and a standard deviation for the pose rmsd of ∼0.05 Å.

Many docking results exhibited "substantially" robust behavior, with robust behavior for most input structures, but with occasional outliers representing somewhat different top-scoring poses. The GlideXP docking to p38 (PDB code: 1ywr) shown in Figure 4 is an example of substantially robust docking behavior. Out of the 500 docked solutions, 16 mild outlier solutions have a similar pose but with the pyrrolidine ring in an axial position. Excluding these outliers allows the remaining data to be described by the normal distribution discussed above. Depending on the number of outliers and how far they are situated from the mean, they can be either considered or excluded from the standard deviation, provided the outlier behavior is properly characterized. The score and rmsd histograms in Figure 4 show minor outlier behavior while the Q−Q plots indicate that a normal distribution describes the majority of score and rmsd points. Using all the data points, the score has a mean and standard deviation of −13.87 ± 0.15 with a pose standard deviation of 0.09 Å. If the outliers are excluded, the pose standard deviation would be 0.07 Å.

The results for GlideHTVS docking to chk1 kinase shown in Figure 5 demonstrate *erratic* docking, characterized by a plethora of docked poses and scores. The rmsd and score histograms show multiple modes and outliers, and the metric scaling plot suggests at least two major (and several minor) binding modes. The Q−Q plots indicate that these distributions are far from being normal, so the mean and standard deviation are not appropriate measures for their description.

Box-Plot Summary of Results across Methods. The rmsd and score variations across all methods and targets are summarized in Figures 6 and 7 using box plots, which report data variations through the sample minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and sample maximum. The spacings between the regions indicate the degree of spread in the data and help identify outlier behavior. If a method has no output variations, its box plot would consist of a single horizontal line. Robust docking results (exemplified by GOLD default docking to cdk2 in Figures 6 and 7) produce box-plots with small Q1−Q3 interquartile boxes and small vertical lines because the majority of solutions are very similar in pose and score. Erratic docking results (exemplified by

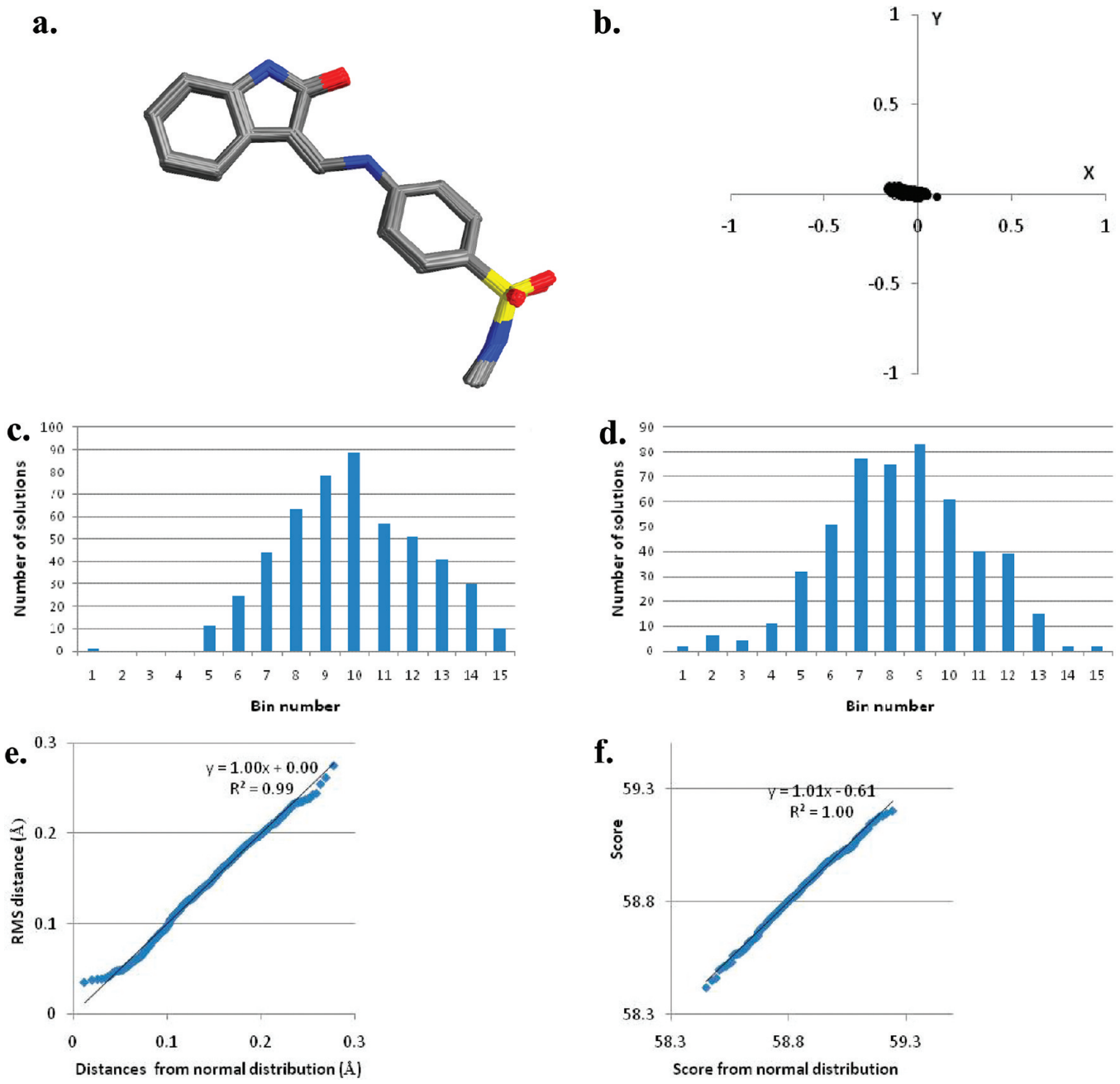


Figure 3. Example of robust docking results, obtained with GOLD using default settings in the cdk2 binding pocket (PDB code: 1ke5). (a) All 500 docking poses overlaid. (b) Metric scaling of the obtained conformations (without coordinate transformation), showing that there is essentially one binding pose/conformation. (c) Histogram showing the distribution of rms distances (measured in place without coordinate transformation, in comparison to the first solution) of the docking poses. The total range covered by the 15 bins is 0.29 Å. (d) Histogram showing the distribution of docking scores. The total score range covered by these 15 bins is 0.79, which is 1.3% of the median score. (e) Q−Q plot showing the fit between the actual pose rmsd and the rmsd calculated from the mean and standard deviation of the data. The high correlation indicates that the distribution is very close to normal. (f) Q−Q plot showing that the scores calculated from the mean and standard deviation fit the actual scores well, i.e. the distribution is normal.

GOLD library screening docking to vegf2) produce box-plots with large interquartile Q1−Q3 boxes and large MAX−MIN ranges for both score and rmsd, reflecting the wide range of scores and binding modes produced by the docking runs.

The methods and/or settings of a docking program are often labeled according to the perceived quality of results. The Glide series distinguishes three methods: the high throughput (HTVS), standard precision (SP), and extra precision (XP) methods. Similarly, the GOLD program has various presets for its genetic algorithm, three of which were studied here: fast library screening (LS), slower default (DF) modes, and its most accurate "very flexible" setting (VF). It is clear from the box plots in Figures 6 and 7 that the fast screening methods—Glide HTVS and GOLD library screening—display erratic behavior for many targets. Exceptions to this behavior are the less open sites (e.g., nuclear receptors) and less flexible ligands (e.g., thymidine kinase), for which even these methods provide relatively robust outcome. In general, the higher quality

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

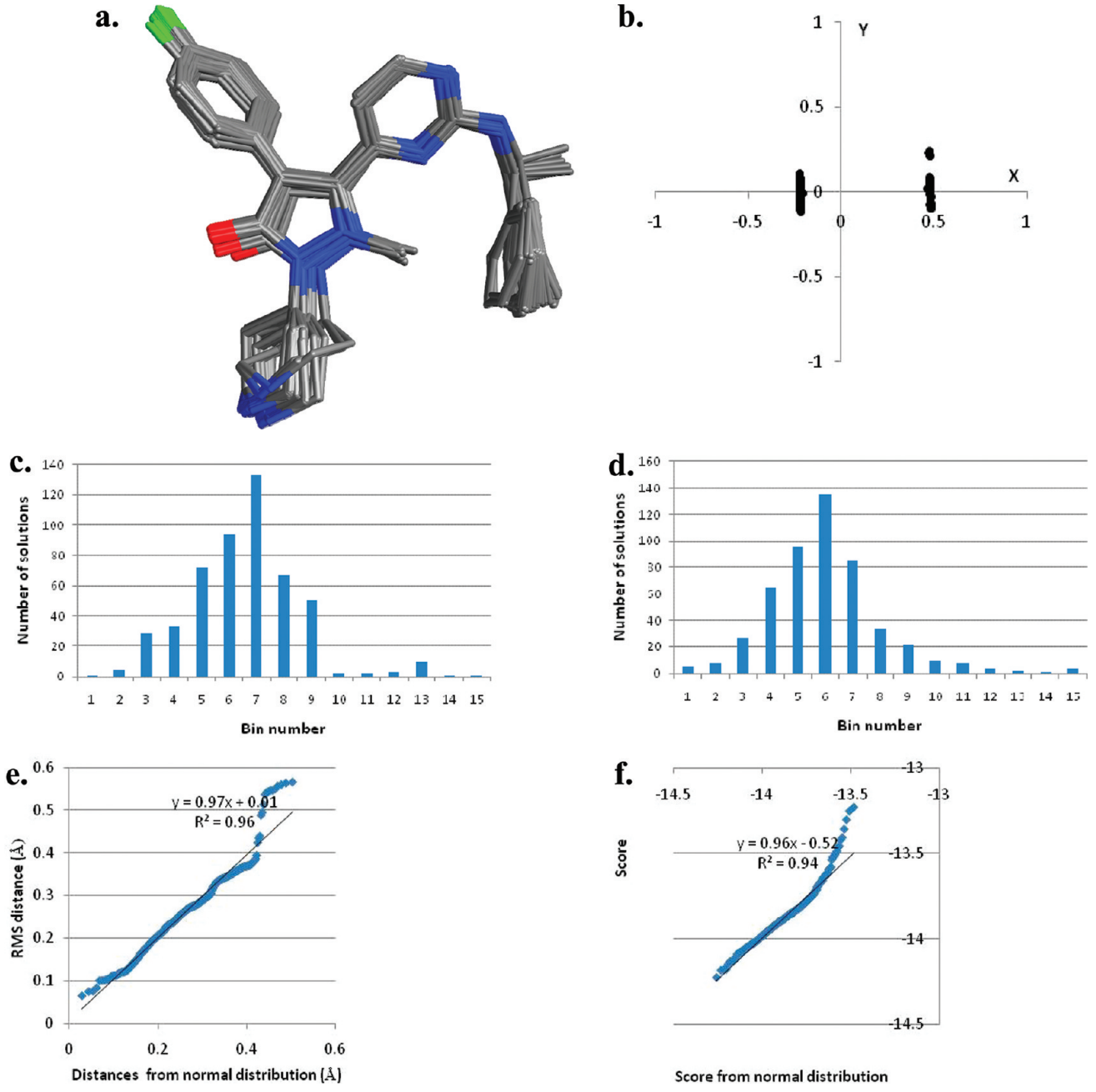


Figure 4. Example of a "substantially" robust docking result, obtained with GlideXP using default settings in the p38 binding pocket (PDB code: 1ywr). (a) All 500 docking poses overlaid, with the majority of solutions in a single conformation and a minority in a slightly different one. (b) Metric scaling of the obtained poses (without coordinate transformation), showing that there were two major binding poses that hardly differ from each other. (c) Histogram showing the rmsd distribution (measured in place without coordinate transformation, in comparison to the first solution) of the docking poses. The total range covered is 0.65 Å. (d) Histogram showing the distribution of docking scores. The total score range covered by these results is 1.0, which is 7.5% of the median score. (e) Q−Q plot showing the correlation between the actual and theoretical pose RMSDs distributions. The plot shows that the distribution is approximately normal with a small number of outliers. (f) Q−Q plot of actual scores versus theoretical scores calculated from the distribution mean and standard deviation. This indicates that the score distribution is approximately normal, with a small number of outliers.

methods (i.e., lower throughput Glide and GOLD settings) provide far more robust behavior, as evidenced by their smaller Q1−Q3 interquartile boxes and MIN−MAX ranges. However, even the higher quality methods occasionally exhibit erratic behavior, as exemplified by GlideXP with chk1 and GOLD default method with p38.

The results in Figures 6 and 7 are summarized in Table 1 as median values for the range and interquartile distances of variations across all targets and methods. The presented values reinforce the above conclusion that high-throughput methods are overall quite erratic, whereas the higher quality methods are substantially more robust. The data in Table 2 summarize the pose rmsd distributions for all targets and methods in this study. When the pose RMSDs follow a normal distribution (arbitrarily defined as having a Q−Q plot with an $R^2 > 0.95$), the table reports the standard deviation (in Å) of the RMSDs and the number of points included in the normal distribution (given in parentheses). The number of outliers is thus 500

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

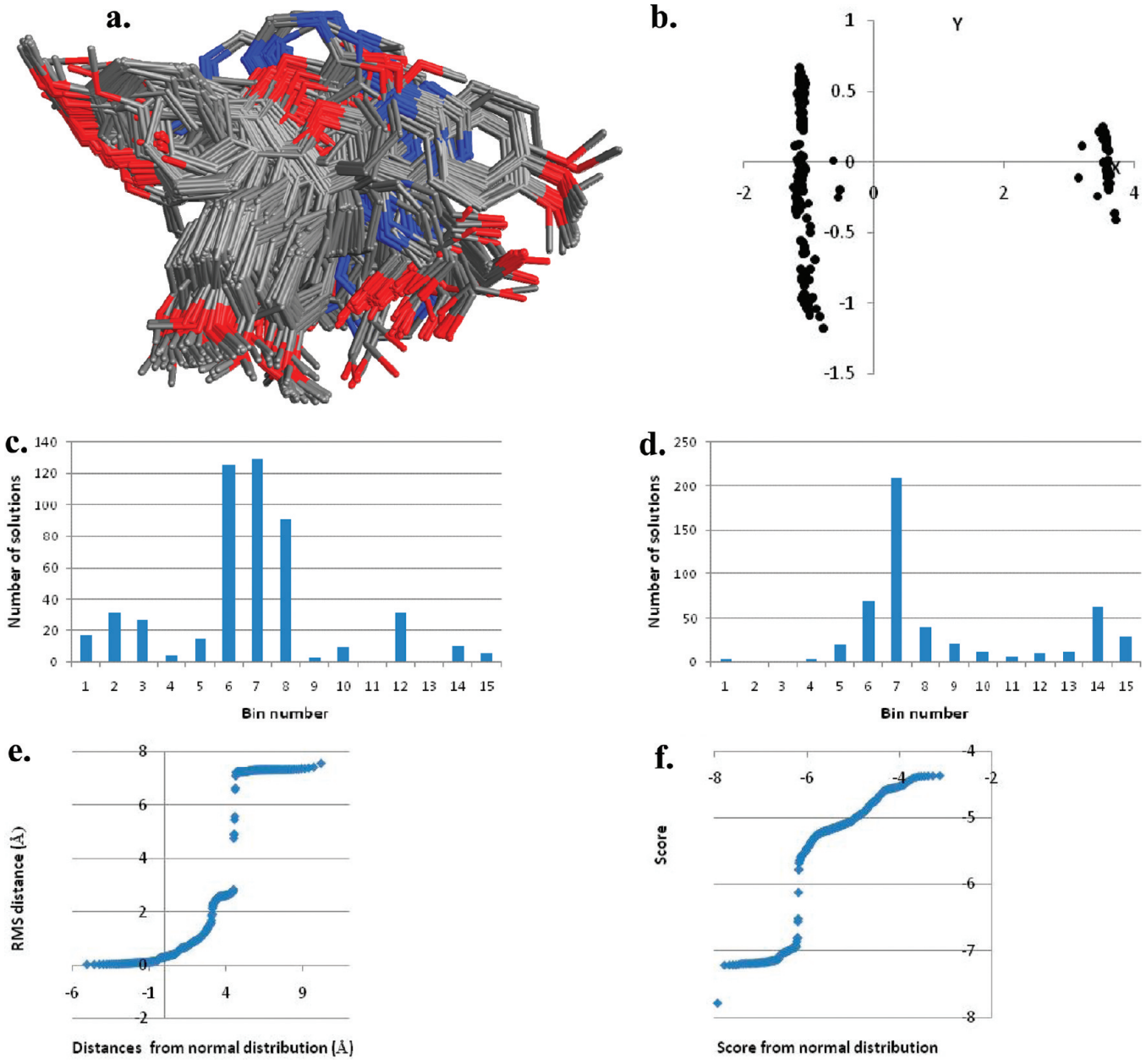


Figure 5. Example of an erratic docking result, obtained with GlideHTVS using default settings in the chk1 binding pocket (PDB code: 2br1). (a) All 500 docking poses overlaid. (b) Metric scaling of the obtained conformations (without coordinate transformation), showing two major binding modes, and a wide range of minor modes within each. (c) Histogram showing the rmsd distribution (measured in place without coordinate transformation, in comparison to the first solution) of the docking poses. The total range covered is 7.6 Å. (d) Histogram showing the distribution of docking scores. The total score range covered by these results is 3.4, which is 62% of the median score. (e) Q−Q plot of the actual rmsd values versus those calculated from the distribution mean and standard deviation which shows low correlation between the two, indicating that the pose rmsd distribution is not normal. (f) Q−Q plot showing low correlation between the actual scores and those calculated from the distribution mean and standard deviation, indicating that the score distribution is not normal.

minus this number. Instances where docking pose distributions are non-Gaussian for more than 20% of the points are indicated with "NG". The results show that many of the Glide results do not follow a Gaussian distribution. The GOLD results seem more likely to be normally distributed, a behavior which may arise from the stochastic nature of GOLD. Even the high-throughput GOLD runs produced normally distributed poses, although the standard deviations are on average 2−3 times greater that the higher precision methods.

Table 1 also summarizes the correlation between the number of rotatable bonds and pose variability, expressed as the Pearson correlation coefficient *R*. As we can see, there is a weak correlation between these quantities for GOLD docking, especially in the very flexible (i.e., the most efficient search) case. For Glide, the correlation with the number of rotatable bonds is less clear, and somewhat counterintuitively, score variability appears to have a small inverse correlation with size.

Relationship between Input Perturbation Size and Docking Result Variations. The question may arise if there is any relationship between the size of the input perturbations and the size of the output pose and score variations. The input structures studied here had maximum torsion angle differences

dx.doi.org/10.1021/ci200598m | *J. Chem. Inf. Model.* 2012, 52, 724−738

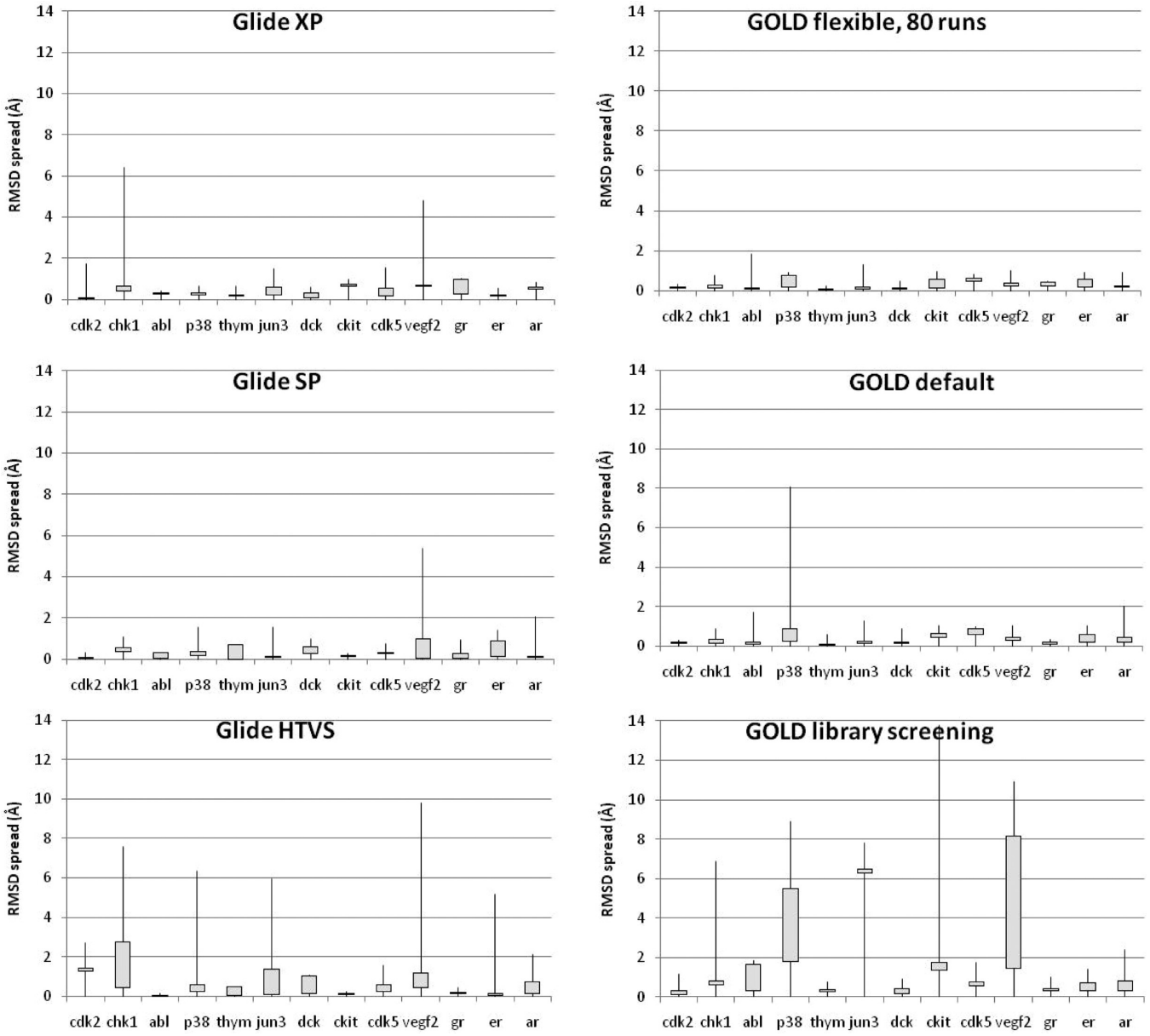


Figure 6. Box plots showing the reproducibility of placement by different docking methods on a selection of kinase and nuclear receptor targets. 500 near-identical copies of the input (the maximum difference in any torsional angle between them was < 0.1°) were docked, and the overlaid solutions were analyzed. The horizontal sides of the boxes represent the first and third quartile of the rmsd to the first solution, whereas the ends of the thin lines represent the rmsd maxima and minima. If these methods had no variation, these box plots would consist of a single horizontal line. See text for further details.

of less than 0.1°, corresponding to a mean RMSD of 0.005 Å between structures. The largest rmsd between any two input conformations is 0.01 Å for the c-kit target, followed by 0.008 Å for the AR target, and 0.007 Å for the ER and p38 targets. On the other end of the scale, the smallest rmsd range was 0.002 Å for the GR target, and 0.003 Å for the CHK1, THYM, JUN3, and DCK targets. Comparing this list with Figures 6 and 7 shows there is generally no correlation between the magnitude of input perturbations and the docking output variations.

To examine effects of input perturbation size on erratic docking results, the Glide HTVS method was chosen to study output variations arising from different sized torsion angle perturbations. In addition to the 500 structure input ensembles with 0.1° torsion angle variations, two additional 500 structure ensembles with larger (1.0°) and smaller (0.01°) maximum torsion angle variations were produced and docked. The pose RMSDs, score ranges, and interquartile distances for the different input sets are displayed in Figure 8. If all of docking output variations were due to some kind of a conformational coverage issue related to input coordinate perturbations, one would expect smaller ligand perturbations to have a proportionately smaller effect on the range of output data. However, the results in Figure 8 show that there is no correlation between output variations and the size of the input torsion perturbations.

Relationship between Starting Structure and Docking Results: (ω_1, ω_2) Torsion Plots. Docking variations between highly similar input structures may simply reflect a cusp region in docking search space, where small input structural changes produce either one docked solution or another. To investigate

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

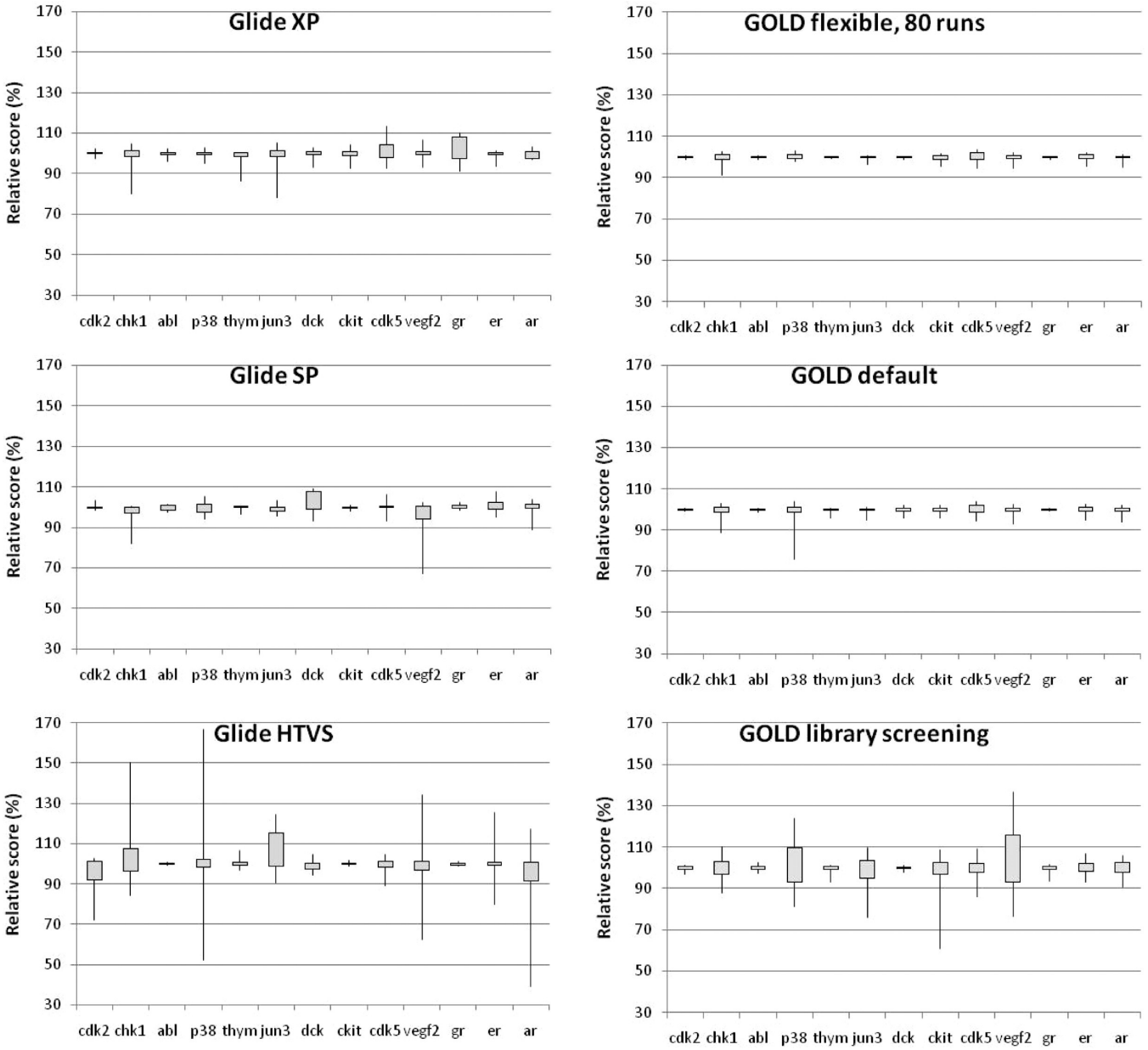


Figure 7. Box plots showing the reproducibility of the calculated docking score by different methods on a selection of kinase and nuclear receptor targets. 500 near-identical copies (the maximum difference in any torsional angle between them was < 0.1°) of the input were docked and the docking scores of the top-scoring solutions were analyzed. In order to reduce the dependence on methods and targets, the scores were rescaled by dividing with the median score and are given as percentages (hence, the median score is always at 100). The horizontal sides of the boxes represent the first and third quartile of the score distribution, whereas the ends of the thin lines represent the maxima and minima of these distributions. If these methods had no variation, these box plots would consist of a single horizontal line. See text for further details.

these regions in detail, input ligand ensembles were created by simultaneously scanning only two torsion angles (ω_1, ω_2) over 1°. The ensemble structures were docked and the top-scoring pose kept in each case. The range of torsion angle variations was small and the number of points large (40 000) in order to get a fine-grained picture of the relationship between closely spaced input conformations and their final docked solution. The results are given in Figures 9−11.

The 1ywr system represents an extreme of docking sensitivity, with some input geometries failing to dock despite the fact that they have near identical input geometries to structures that dock successfully. This pass/fail behavior might arise from relatively snug fits, as is the case for the 1ywr structure. The torsion scan of this system docked with GlideHTVS in Figure 9 shows the pass/fail state of the docking run (i.e., docking succeeded or failed to produce a docked solution) versus the (ω_1, ω_2) torsion angles of the input structure. It is interesting to note that the failure cases do not form a single contiguous region in the plot and there are no well-defined boundaries between regions where (ω_1, ω_2) combinations either succeed or fail to dock. However, there does appear to be a faint repeating sawtooth pattern in the plot, suggesting some nonrandom systematic effect is playing a role in determining the pass/fail state of the docking run.

Unlike the 1ywr system, most systems considered in this study always produce a docked pose when highly similar ligand

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

Table 1. Assessment of the Robustness of Docking Methods and Settings over Different Targets[a]

	pose range (Å)[b]	pose interquartile distance (Å)[b]	score range[c]	score interquartile distance[c]	R (pose variability and N_{rot})[d]	R (score variability and N_{rot})[e]
Glide XP	1.0	0.1	11.3	1.6	0.43	−0.19
Glide SP	1.0	0.2	11.3	2.1	0.37	−0.50
Glide HTVS	2.1	0.4	30.5	3.1	0.35	−0.49
GOLD flexible 80	0.9	0.1	5.2	0.8	0.71	0.64
GOLD default	1.0	0.1	6.4	1.6	0.64	0.38
GOLD library	1.8	0.2	15.3	4.3	0.47	0.51

[a]Robustness was assessed by docking 500 nearly identical copies of the input (the maximum difference in any torsional angle between them was <0.1°) and assessing the output in terms of placement and score. Median values across 13 targets are presented throughout this table. See text for further details. [b]Range and interquartile distance were obtained by calculating in-place rms distances of each pose to the first solution. From the obtained distribution of rmsd values, the largest value corresponded to the range and the difference between the values representing the first and third quartile provided the interquartile distance. Ranges and interquartile distances were calculated for each target (see Figure 6), and the medians of these values are provided in this table. [c]Score ranges and interquartile distances were obtained by considering the scores of the 500 docked solutions. Since the scores vary strongly for different targets and in different methods, they were scaled by dividing with the median score. Ranges (the absolute value of the difference between the highest and lowest score) and interquartile distances (the absolute value between the scores corresponding to the first and third quartile) were calculated for each target (see Figure 7), and the medians of these values are provided in this table. [d]The correlation coefficient (R) between the pose variability (as expressed by the mean rmsd range of the poses) and the number of rotatable bonds (b_rotN in MOE). [e]The correlation coefficient (R) between the score variability (as expressed by the range of score values over the median) and the number of rotatable bonds (b_rotN in MOE).

Table 2. Analysis of Docking Solutions with Approximately Normal Distribution: Standard Deviations of the pose RMSD and the Number of Data Points That Follow This Distribution (in Brackets)[a]

PDB code and target[b]	Glide HTVS	GlideSP	GlideXP	GOLD library	GOLD default	GOLD 80	GOLD flexible 80
1ke5_cdk2	NG	NG	NG	0.106 (461)	0.048 (500)	0.040 (500)	0.028 (498)
2br1_chk1	NG	NG	0.162 (498)	0.148 (497)	0.113 (496)	0.106 (496)	0.103 (498)
1opk_abl	NG	NG	NG	NG	0.041 (500)	0.047 (500)	0.134 (495)
1ywr_p38	NG[c]	0.135 (497)	0.071 (484)	NG	NG	NG	NG
1of1_thym	NG	NG	NG	0.110 (500)	0.020 (456)	0.030 (474)	0.016 (489)
1pmn_jun3	NG	0.018 (433)	0.237 (492)	NG	0.044 (471)	0.066 (495)	0.063 (488)
1p62_dck	NG	NG	NG	0.198 (500)	0.047 (450)	0.058 (481)	0.031 (488)
1t46_ckit	0.036 (500)	0.013 (487)	NG	0.244 (497)	0.172 (500)	NG	NG
1unl_cdk5	NG	NG	NG	0.141 (487)	NG	NG	NG
1y6b_vegf2	NG	NG	NG	NG	0.083 (411)	0.091 (412)	0.084 (439)
1m2z_gr	NG	NG	NG	0.125 (498)	0.069 (500)	NG	NG
1sj0_er	NG	NG	NG	0.236 (500)	NG	NG	NG
1z95_ar	NG[d]	NG	0.055 (480)	0.275 (488)	0.121 (433)	0.129 (493)	0.058 (478)
median			**0.12**	**0.15**	**0.06**	**0.06**	**0.06**

[a]500 highly similar inputs (with maximum difference between any torsional angle < 0.1°) were docked, the rmsd between the first and all other docked solutions were calculated in place, plotted using a Q−Q plot, and sections approximating normal distribution were characterized. The criteria were that the selected points had to be continuously distributed on the plot, they had to be approximated well by values calculated from Gaussian distribution (r^2 > 0.95 and a slope between 0.95 and 1.05), and at least 80% of the points (>400 points) had to be included in this section. NG = non-Gaussian distributions. [b]The PDB code and abbreviation of the target. See text for more details. [c]There was a section with an almost perfectly Gaussian distribution involving 315 points and a standard deviation of 0.051; the rest of the solutions belonged to multiple modes covering a large range of poses and scores. [d]Only produced 357 solutions, the other 143 docking attempts failed.

inputs were used, even if the range of top-scoring poses produced is large. One approach to understand the outcome in these cases is to cluster the docked poses and scores, as was done for the 40 000 top scoring poses from a ztorsion scan of the 2brk_chk1 system. On the basis of absolute coordinates, the poses were clustered into six pose clusters as described in the Methods section. A detailed description of the pose clusters and their score distributions are given in the Supporting Information. The largest cluster by far was pose cluster 1, with 36 171 of the 40 000 starting structures producing a top-scoring pose in pose cluster 1. A significant amount of input structures (3141) produce a top-scoring pose in pose cluster six. The remaining clusters represent small percentages of the total results. In Figure 10, the pose cluster membership of the top pose is plotted as a function of the input torsion angles (ω_1, ω_2). The plot shows large contiguous regions where most (ω_1, ω_2) torsion combinations produce a top pose in cluster 1, and well-defined boundaries between regions where (ω_1, ω_2) combinations produce poses in either cluster 1 or cluster 6. The plot also contains erratic regions where small (ω_1, ω_2) variations converge to a number of different pose clusters, despite tiny differences in the input torsion angles.

The docking scores for the torsion scan of the 2brk_chk1 system were binned and plotted versus the starting torsion angle (ω_1, ω_2) in Figure 11. The bins were originally all set to 0.5 score units starting at a score of −8.0, but because a large portion of cluster 1 poses had scores near the −7.5/−7.0 bin boundary, the two bins between −7.5 and −6.5 were combined into a single bin. The score bin plot is quite similar to the pose cluster plot, with contiguous (ω_1, ω_2) regions where scores fall

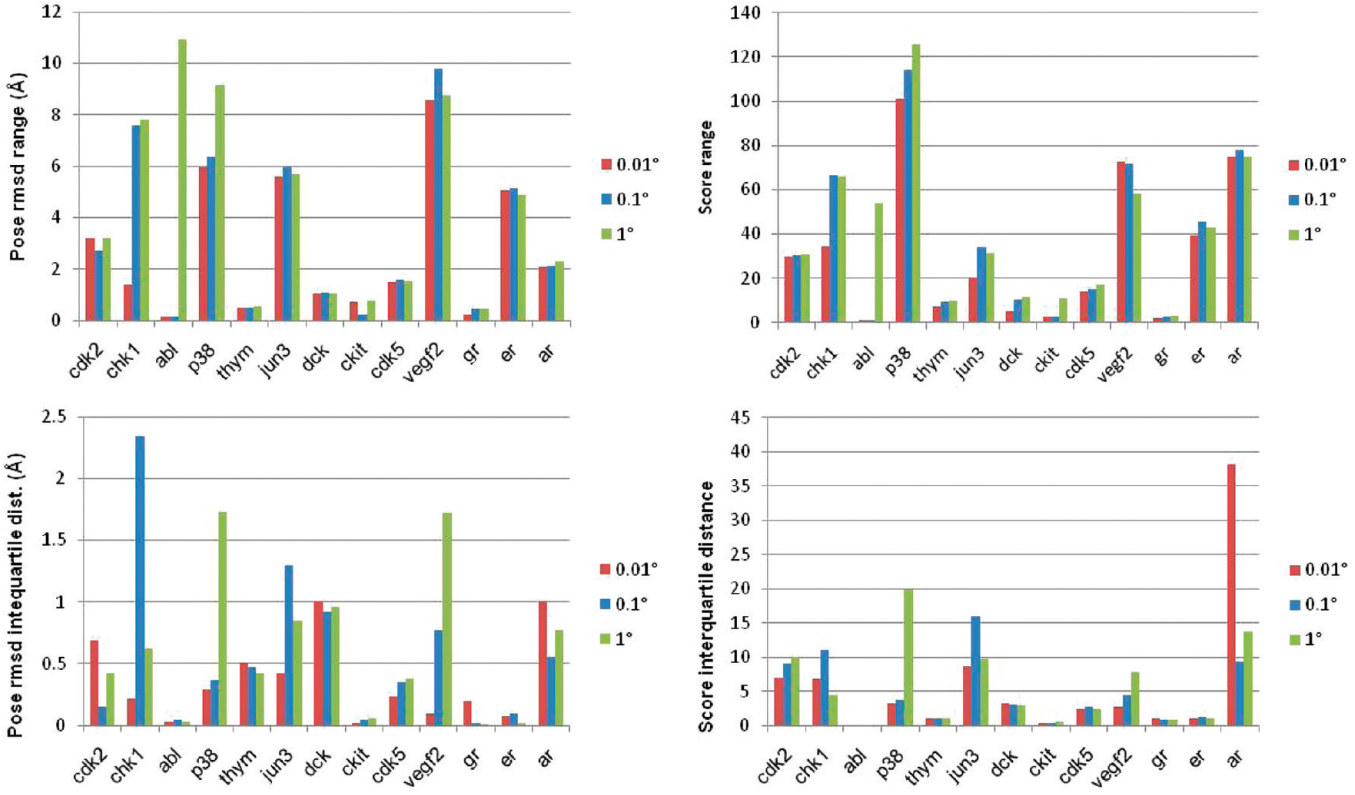


Figure 8. A comparison of docking solution distributions from high throughput Glide docking (Glide HTVS) for 13 kinase and nuclear receptor targets. The 500 input structures in each case were highly similar to each other, with a maximum difference in any torsional angle being less than 1°, 0.1°, and 0.01°, respectively. The bar graphs show the largest rmsd difference from the first structure (which corresponds to the range of rmsd values), the range in relative scores, as well as the corresponding interquartile distances. If output variations arose strictly from the coordinate differences and not from the numerical sensitivity of the program, we would expect to see gradual improvement in the reproducibility of docking results when decreasing the difference between the input structures. This, however, is not the case, suggesting numerical sensitivity is in part responsible for the output variations.

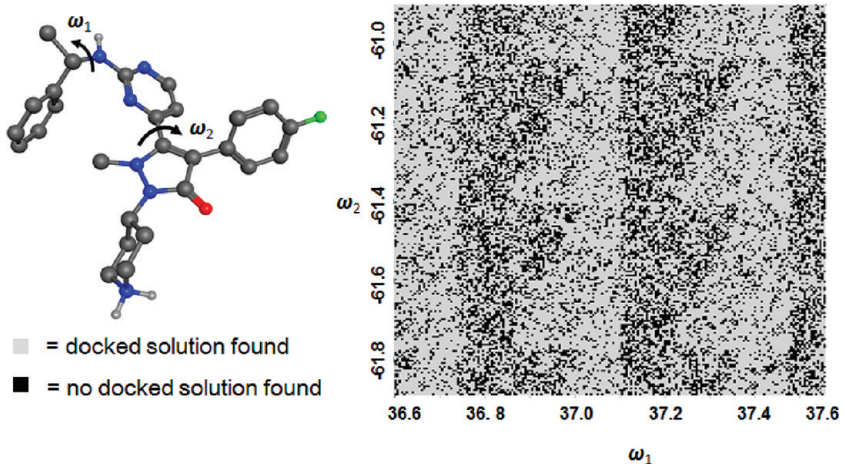


Figure 9. Pass/fail state of docking the cognate ligand to the 1ywr structure as a function of two torsion angles. 40 000 input structures were generated around the original Corina structure by varying ω_1 and ω_2 as indicated in the figure by a maximum of 1°. Input combinations of (ω_1, ω_2) which produce a docked solution are plotted as gray points, while (ω_1, ω_2) combination that failed to produce any docked poses are plotted as black points.

into one bin, well-defined (ω_1, ω_2) boundary regions between score bins, and erratic regions where small (ω_1, ω_2) changes can give rise to a number of possible scores. As with the 1ywr torsion scan plot, the pose cluster and score torsion plots for the 2brk_chk1 system exhibits repeating but nonregular patterns in the data, suggesting a nonrandom effect is the cause of the variations.

It should also be noted that the (ω_1, ω_2) torsion plots in Figures 9−11 have features similar to mappings of systems that exhibit *final state sensitivity*[16]—a condition characterized by *fractal boundaries* between regions of input space that optimize to different final states. As discussed at length by Yorke et al.[16] in the context of simple physical systems, *basin boundaries* are lines that separate regions of input space which optimize to

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

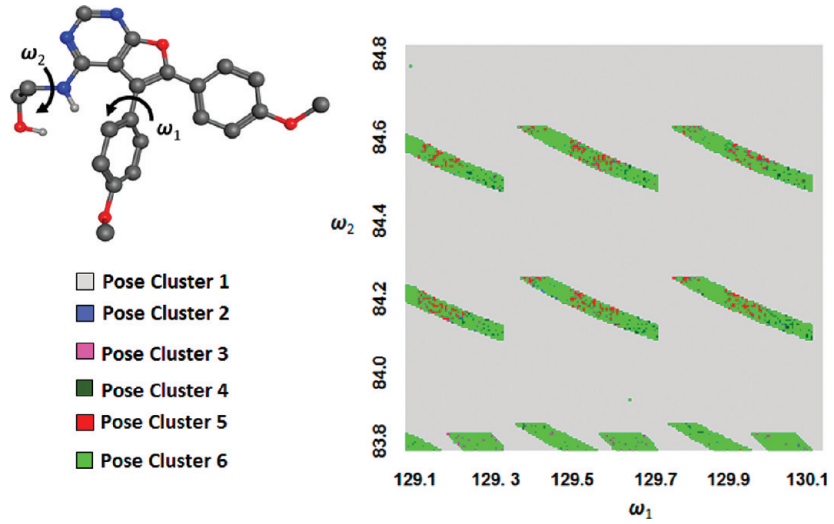


Figure 10. Pose cluster membership of the top-scoring pose of the 2br1 ligand as a function of two torsion angles. 40 000 input structures of the cognate ligand were generated around the original Corina structure by varying ω_1 and ω_2 as indicated in the figure by a maximum of 1°. See text for details.

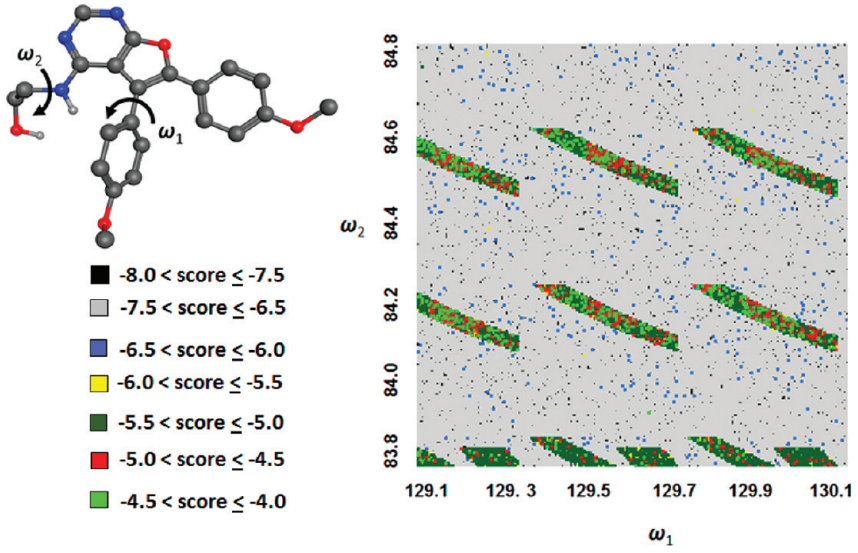


Figure 11. Score bin membership of the top-scoring pose of the 2br1 ligand plotted versus two torsion angles. 40 000 input structures of the cognate ligand were generated around the original Corina structure by varying ω_1 and ω_2 as indicated in the figure by a maximum of 1°. See text for details.

different final states. In mathematical terms a set of input values which all converge to the same unique minimum are called the *attraction basin* of that minimum. One can investigate the nature of basin boundaries by plotting the final state of a system as a function of varying input parameters (akin to the plots in Figures 9−11), which produces a map that shows both the shapes of the attraction basins and the nature of the boundaries between them. In nonchaotic systems, boundaries between attraction basins are smooth continuous lines that clearly separate input space into regions which optimize to different minima. In chaotic systems, boundaries between attraction basins can form irregular, fractal-like curves. When initial conditions are near fractal boundaries, proximal starting points can converge to distant final minima. In chaotic systems, plotting the final state as a function of starting points produces attraction basins with intricate, disconnected boundaries instead of contiguous regions with smooth, well-defined boundaries. Well-known examples of fractal basin boundaries include the logistic difference equation and Newton's method applied to solving complex roots.[16] Fractal boundaries have already been reported[17] in MM minimizations of small molecules, so their presence in docking searches is not inconceivable.

Effect of Atom Shuffling on Docking Results. One may argue that docking variations observed as a result of ligand coordinate perturbations are expected because the input structures have different (albeit almost identical) Cartesian coordinates, and are thus "different" from a scoring function and energy calculation perspective. Shuffling (or *permuting*) the order of the atoms in a ligand input file is a method of introducing differences into ligand input files *without* changing atom coordinates. In the atom-shuffled (or *atom-permuted*) files the Cartesian coordinates and other attributes of the ligand remain unchanged, so the structures should be identical from a scoring function and energetics point of view; the only difference is the order in which the atoms are specified in the file. In practice, atom order in an input file is determined by a

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

number of factors, but usually reflects the order in which the atoms of a molecule were either drawn in a sketching package, generated from a SMILES string, or created in a 3D builder.

The effect of atom shuffling on docking results was investigated by creating 500 *permuted* ligand input files (which differed only by the order of the atoms in the file), followed by docking to the respective targets and retaining the top-scoring pose. The resulting variations in top-scoring pose for the 1y6b structure are plotted in Figure 12, along with

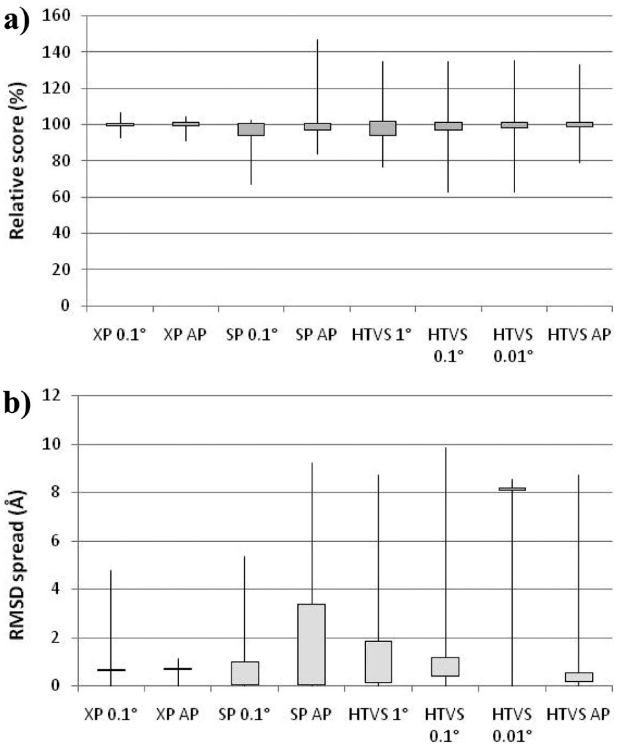


Figure 12. Box plots showing the variability in pose and score using different input structures and levels of precision with Glide for the 1y6b structure (vegf2): (a) Relative scores and (b) rms spread of the poses. Each bar graph represents 500 structures of the cognate ligand that either had coordinate perturbations (all torsional angles varied by a maximum of 0.01°, 0.1°, or 1° as indicated on the axis) or the order of their atoms were permuted (leaving the actual coordinates intact, marked as AP) using GlideXP, GlideSP, and GlideHTVS (marked as XP, SP, and HTVS, respectively). The horizontal sides of the boxes represent the first and third quartile of the rmsd and the thin lines represent the rmsd maxima and minima. The surprising finding is that although the atom permuted inputs differ only by the order in which the atoms are listed in the file (the Cartesian coordinates and other ligand attributes are all identical), the docked structures still have substantial variability in both pose and score, comparable to inputs with small coordinate perturbations of various magnitude. See further details in the text.

variations from coordinate-perturbed inputs for comparison. If docking variations arose solely from ligand coordinate perturbations, the atom-permuted input file would produce no variations in the final results. However, it is clear from Figure 12 that simply permuting the atom order in ligand input files can lead to variations in docking results, in some cases larger than those produced from coordinate perturbations. This fact demonstrates that numerical sensitivity in docking algorithms is also a factor in docking variability. Since digital calculations produce small variations depending on the order in which operations are performed,[1] shuffling the atoms in an input file can affect docking because the atom order can determine the summation order of MM energy and scoring function terms. This could potentially lead to small variations between input structures—despite the fact they have identical Cartesian coordinates. These small variations get magnified by the docking algorithms and grow into significant variations in the docking output. Although only results for the Glide software package are presented here, we have observed atom order sensitivity in other commercial deterministic docking programs and will report those findings in future publications.

CONCLUSIONS

This study clearly demonstrates that seemingly insignificant differences in ligand input, such as small coordinate perturbations or permuting the atom order in an input file, can have a dramatic effect on the final top-scoring docked pose. Although adding small coordinate perturbations or permuting the atom order in an input file may seem contrived, it reflects many real-world scenarios where small but unperceivable differences can be introduced into ligand input structures, even when efforts are made to minimize these differences. Input prepared by different individuals, or even by the same person at different times, can show variations due to the source of the molecule (SMILES string, SD file, etc.), the order in which the ligand atoms are drawn, the input conformation generation method (e.g., Corina, ligprep, conformational search) and the geometry optimization method used to refine the structure (force fields applied, solvation, cut-offs, etc.). Even if the structure is prepared in exactly the same way, small changes can be introduced in the last digits of the structure coordinates during molecule manipulations and file input−output. In practice, this means docking results can be affected by how one prepares the input structure—which could lead to confusion for any unsuspecting user who repeats a docking experiment with *almost exactly* the same procedure, but obtains wildly different results. In some cases docking results may follow well-behaved normal distributions, but in many cases the results can be erratic, with a large range of solutions representing multiple binding modes. This suggests that all docking studies should, at least in principle, consider output variability, with statistical analysis of pose/score distributions, proper outlier characterization and error bars.

It is important to note that a docking program with lower sensitivity to input is not necessarily a better program, the quality needs to be established with reference to experimental data. However we believe that this variability is an important and previously neglected factor in docking quality. The output variations of the high-throughput methods (GlideHTVS and Gold Library Screening) are large and highly dependent upon the input structure, bringing into question their ultimate usefulness. The more precise docking settings showed much less variability overall, but large variations and extreme outliers were occasionally observed with these methods as well, indicating caution should be exercised in their use. One can imagine that if the few observed outliers were the only inputs used in a docking study, profoundly different conclusions about poses and ranking would result. However following our previous recommendations,[6] the effect of such outliers can be reduced by using multiple input exemplars. This method might significantly slow down calculations but future improvements in algorithms, CPU speed, and leveraging gpu speed will make the approach increasingly practical.

dx.doi.org/10.1021/ci200598m | J. Chem. Inf. Model. 2012, 52, 724−738

The sensitivity of docking to input perturbations and atom permutations supports our previous findings which suggested that part docking sensitivity displays the hallmarks of chaotic behavior[18,19] as observed in meteorological,[18] seismological,[20] and celestial trajectory simulations.[21] In these and other systems, which are *sensitive to initial conditions*, small changes to initial conditions can lead to divergent trajectories and conflicting conclusions. The chaotic nature of molecular dynamics (MD) trajectories is well-documented,[22−25] and often manifests itself as a drift in the total energy. In many cases, the chaotic nature of a simulation can be properly characterized with theoretical measures (such as Lyapunov exponents[18] for diverging trajectories), but quantifying chaotic effects in docking is not so straightforward; since docking is a combination of algorithms and procedures, a full characterization would examine each algorithm individually to assess its contribution to the overall numerical sensitivity and determine if it exhibits chaotic regimes. Using the (ω_1, ω_2) torsion plots and similar constructs to characterize chaotic behavior in the individual components of docking algorithms will be the subject of future reports.

Finally, although it may be difficult to prove whether or not the behavior of docking programs is indeed chaotic, the main practical issue remains unchanged; docking calculations are highly sensitive to initial conditions, and conflicting results can be produced from near-identical starting points. This issue of reproducibility (or lack thereof) in advanced simulations goes beyond computational chemistry and docking, and is a general phenomenon in many computational fields. This was addressed in a recent publication from Diethelm,[26] who states

> *"The mathematical background of non-reproducibility is simple and well known within the scientific computing community. Because it is so ubiquitous in numerical high performance algorithms, one has learned to live with it and to draw the right conclusions from the variations in the output data. Thus, since the problem is so well understood in this peer group, a common assumption is that it is more of a feature than a problem, and one may conclude that reproducibility does not seem to be a requirement that absolutely must be enforced."*

What this means in practice is that although improved algorithms might be devised to reduce this effect, numerical sensitivity in docking calculations is probably here to stay. Note that in this work we implicitly assumed that the target structure is constant in these calculations; in reality it is also variable (variability is introduced during the crystallographic refinement and subsequent structure manipulations, as well as grid generation) and is expected to introduce even more variability in docking output. Hopefully, future algorithms will reduce the deleterious effects of numerical sensitivity arising from these and other sources, so that we can draw the right conclusions from docking calculations.

■ ASSOCIATED CONTENT

Supporting Information

Representative subset of each pose cluster, as well as a histogram of score distributions for these clusters are provided for the 2br1 example, corresponding to the results presented in Figures 10 and 11. This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +1 416 581 7611. E-mail: mfeher@uhnres.utoronto.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Martin Santavy of Chemical Computing Group for discussions about the effect of the order of operations on the reproducibility of digital calculations.

■ REFERENCES

(1) He, Y.; Ding, C. H. Q. Using Accurate Arithmetics to Improve Numerical Reproducibility and Stability in Parallel Applications. *J. Supercomput.* **2001**, *18*, 259−277.

(2) Williams, C. I.; Feher, M. The effect of numerical error on the reproducibility of molecular geometry optimizations. *J. Comput-Aided Mol. Des.* **2008**, *22*, 39−51.

(3) Braxenthaler, M.; Unger, R.; Auerbach, D.; Given, J. A.; Moult, J. Chaos in Protein Dynamics. *PROTEINS: Struct. Funct. Genet.* **1997**, *29*, 417−425.

(4) Feher, M; Williams, C. I. Effect of Input Differences on the Results of Docking Calculations. *J. Chem. Inf. Model.* **2009**, *49*, 1704−1714.

(5) Onodera, K.; Satou, K.; Hirota, H. Evaluations of Molecular Docking Programs for Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1609−1618.

(6) Feher, M.; Williams, C. I. Reducing Docking Score Variations Arising from Input Differences. *J. Chem. Inf. Model.* **2010**, *50*, 1549−1560.

(7) *GOLD*, version 4.1; Cambridge Crystallographic Database: Cambridge, U.K., 2009.

(8) *Glide*, version 2010; Schrodinger Inc.: Portland, OR, USA, 2010.

(9) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(10) *Corina*, version 3.2; Molecular Networks GmbH: Erlangen, Germany, 2006.

(11) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1996**, *17*, 490−641.

(12) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1999**, *20*, 720−741.

(13) Unpublished modification to MMFF94s; enforces planarity of conjugated nitrogens.

(14) Feher, M.; Schmidt, J. M. Identifying Potential Binding Modes and Explaining Partitioning Behaviour Using Flexible Alignments and Multidimensional Scaling. *J. Comput-Aided Mol. Des.* **2001**, *15*, 1065−1083.

(15) Feher, M.; Schmidt, J. M. Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810−818.

(16) McDonald, S. W.; Gregbogi, C.; Ott, E.; Yorke, J. A. Fractal Basin Boundries. *Phys. D* **1985**, *17*, 125−153.

(17) Xu, Y. Z.; Ouyang, Q.; Wu, J. G.; Yorke, J. A.; Xu, G. X.; Xu, D. F.; Soloway, R. D.; Ren, J. Q. Using fractal to solve the multiple minima problem in molecular mechanics calculation. *J. Comput. Chem.* **2000**, *21*, 1101−1108.

(18) Lorenz, E. *The Essence of Chaos*; University of Washington Press: Seattle, 1996; ISBN 0295975148.

(19) Gleick, J. *Chaos: Making a new Science*; Penguin Books: New York, 1987; ISBN 0140092501.

(20) Sornette, D. *Nature Debates: Earthquakes*; 1999; http://www.nature.com/nature/debates/earthquake/ (accessed 9/19/2002).

(21) Groison, D. Est-il vrai que les ordinateurs font des erreurs de calcul ? *Sci. Vie* **2002**, *1022*, 130.

(22) Barth, E.; Schlick, T. Extrapolation versus impulse in multiple-time stepping schemes. II. Linear analysis and applications to Newtonian and Langevin dynamics. *J. Chem. Phys.* **1998**, *109*, 1633−1642.
(23) Biesiadecki, J. J.; Skeel, R. D. Dangers of Multiple Time Step Methods. *J. Comput. Phys.* **1993**, *109*, 318−328.
(24) Bishop, T. C.; Skeel, R. D.; Schulten, K. Difficulties with multiple time stepping and fast multipole algorithm in molecular dynamics. *J. Comput. Chem.* **1997**, *18*, 1785−1791.
(25) Sandu1, A.; Schlick, T. Masking Resonance Artifacts in Force-Splitting Methods for Biomolecular Simulations by Extrapolative Langevin Dynamics. *J. Comput. Phys.* **1999**, *151*, 74−113.
(26) Diethelm, K. The Limits of Reproducibility in Numerical Simulation. *Comp. Sci. Eng.* **2012**, *14*, 64−72.