Reducing Docking Score Variations Arising from Input Differences

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

Received May 21, 2010

The variability of docking results as a function of variations in ligand input conformations was studied for the GOLD, Glide, FlexX, and Surflex programs. It is concluded that there are two major effects leading to such variability: the adequacy of conformational search during docking and random "chaotic" effects arising from sensitivity to small input perturbations. It is shown that although the former is generally the stronger effect, the latter is also highly significant for almost all docking engines. The strong target-to-target variation of the magnitude of these effects is emphasized. The performance of different packages is compared using these measures. Guidelines are provided for different programs to reduce variability and improve reproducibility, which involve using a small number of input conformations as starting points for docking, followed by the selection of the top scoring docked pose from the results as the best docked solution.

INTRODUCTION

One of the most important goals of ligand docking and scoring is to rank molecules and predict their relative binding affinities to the target of interest. Various examples in the literature demonstrate that it is often a daunting task with little success.[1−4] As described in these and other studies, shortcomings in the predictive ability of docking include target-related issues such as target flexibility,[5] inaccurate estimation of ligand strain energy,[6,7] and the general inaccuracy of scoring functions at predicting binding free energy.[8] More recently, experimental artifacts in protein X-ray structures have been shown to contribute to problems with training and validation of docking programs.[9,10] As a result of these issues, it is often difficult if not impossible to gain any meaningful correlation between docking scores and binding affinity for a diverse set of ligands.[11]

We have recently added sensitivity of docking results to variations in ligand input geometry as another potential factor contributing to the poor performance of docking calculations.[12] Docking studies typically produce a list of docked poses and scores starting with a single low-energy conformation of the ligand as input. However, by using ensembles of the same molecule as input to different docking runs, we found that the lists of docked poses and scores varied greatly with ligand input geometry. The final best scores often covered several orders of magnitude in predicted binding affinity, and in many cases, the predicted binding mode of the docked ligand differed with different ligand input geometries. This sensitivity of docking output to ligand input geometry was exhibited by all the programs studied, including those with deterministic algorithms. Interestingly, the smallest variability in results was displayed by the GOLD program, which uses an inherently stochastic algorithm.

In order to better understand our previous and current experimental findings, we carried out some "thought experiments", looking at various possible docking scenarios. Our objective was to understand how variations in the input might lead to variations in the output. The docking process is generally described as being made up of two steps: conformational exploration of the ligand and placement into the active site, followed by scoring that is supposed to assess the quality of fit into the active site. In this work we generally start with a set of ligand conformations generated prior to docking (referred to as *input conformation space*). Conceptually we can describe the end result of docking as a set of corresponding top-scoring docked "solutions" that consist of placements and scores, which will henceforth be referred to as *docked solution space*. As far as the docking engines are concerned, the ensembles of input conformations all represent the same starting point, since they are all valid conformations. We can then imagine various scenarios where variations in the ligand input geometry may affect the conformations produced during docking and how this could influence the distribution of the final docked solutions. (Although it is understood that some docking programs do not explicitly search ligand conformational space, the underlying principles remain the same.) Three such docking scenarios are depicted in Figure 1; these can be thought of as limiting cases. In Figure 1, ligand input conformation space and docked solution space are represented as closed regions with individual conformers and docking solutions represented as points (stars, circles, diamonds) in each region. The distance between points in ligand input conformation space is proportional to the dissimilarity between the input conformers (proximal points = similar conformations). The distance between points in docked solution space reflects the similarity of the docked solution in both pose and score. In the docking solution space, only solutions with similar poses *and* scores are represented as proximal points. Docked

* Corresponding author phone: +1 416 581 7611; e-mail: mfeher@uhnres.utoronto.ca.

† Campbell Family Institute for Breast Cancer Research.

‡ Chemical Computing Group.

10.1021/ci100204x
Published on Web 08/10/2010

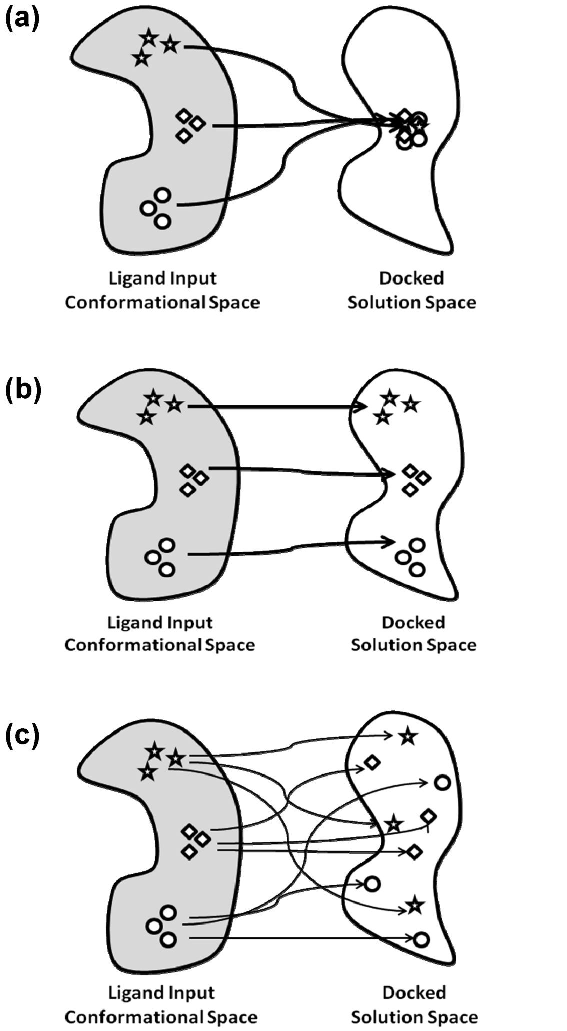
(a)
Ligand Input Conformational Space
Docked Solution Space
(b)
Ligand Input Conformational Space
Docked Solution Space
(c)
Ligand Input Conformational Space
Docked Solution Space

Figure 1. Relationships between ligand input conformation space and docked solution space. (a) The "ideal" docking scenario. Differences in ligand input geometry have little or no effect on the final results. The same best scoring docked solution (pose and score) is produced regardless of ligand input conformation (assuming the ligand is in a reasonable conformation). (b) Conformational coverage is the main issue driving docking result differences. Due to lack of coverage in the docking conformational search, certain input conformations can never access certain docked poses. However, similar input conformations are expected to produce, within a range, similar docked solutions. (c) Docking shows sensitivity to initial conditions. Here, small differences between input conformations are magnified by the docking process, resulting in a divergence of their simulation trajectories. Similar ligand input geometries can then produce vastly different docked solutions with large variations in pose and/or score.

solutions that differ in score but are similar in pose (or vice versa) are more distant.

In an ideal docking scenario (Figure 1a) there exists a single "global" minimum docked solution, and the conformational and docking pose search routines are both sufficiently complete that all reasonable input ligand conformations (represented by stars, circles, and triangles) converge to the same global minimum docked solution. The results here will be independent of ligand input geometry. This scenario is what is assumed by many docking studies and virtual screening protocols. The diagram in Figure 1b represents a second scenario, where the search of ligand conformational space is incomplete, making certain docked poses inaccessible to certain input conformations. However, if there are no "random" or stochastic elements in this process, similar input conformations will converge to similar docked solutions, i.e, the local minimum docked solution for that range of input conformations. On the other hand, dissimilar input conformations can converge to different docked solutions. Behavior of this sort could conceivably be remedied by running a set of diverse ligand geometries as input to overcome deficiencies in the docking program conformational search routine. The global minimum docked solution could then be found as the best scoring pose from all of the input ligand runs. Finally, Figure 1c represents a docking scenario that exhibits *sensitivity to initial conditions*, where small variations in ligand input conformations produce wildly different final poses and scores. In general, simulations which are sensitive to initial conditions (sometimes also referred to as "*chaotic*")[13,14] exhibit instabilities when faced with small input perturbations, as each iterative cycle of the algorithm accumulates and magnifies small input differences to a point where the simulation trajectories diverge substantially. Well-known examples of such systems include meteorological,[13] seismological,[15] and celestial trajectory simulations.[16] Sensitivity of molecular dynamics (MD) simulations to initial conditions and the resulting trajectory instabilities are well-documented.[17−20] Even molecular mechanics minimizations of large systems have been shown to exhibit initial condition sensitivity.[21] Docking routines may exhibit sensitivity to initial conditions due to an incomplete sampling of ligand conformation space, an incomplete sampling of pose space, and/or a scoring function with many deep local minima and regions with steep gradients that would be sensitive to small ligand pose perturbations. Large variations in docking scores can arise from small changes in sensitive scoring function terms, such as H-bond distances, changes in numerous small hydrophobic interactions, and in some cases large pose changes. Although a complete dissection of the scoring function terms and their relative effects on score variations is beyond the scope of this work, the sensitivity of each scoring function term to input variability should be considered by any future workers developing scoring functions.

These effects are depicted in Figure 1c; here, similar input conformations may not converge to the same docked solution, while dissimilar input structures may converge to the same docked solution. Furthermore, similar poses may show large score variations while dissimilar poses may exhibit similar scores. Overall, docking routines that are sensitive to initial conditions could display what appears to be random behavior when subjected to small perturbations in the ligand input structure.

Clearly, if docking is to produce meaningful rankings, the behavior of docking simulations with respect to variations in ligand input geometry needs to be investigated in a meaningful way. In this paper, we attempted to obtain a better understanding of the nature of this variability, with the aim of developing approaches to reduce its effects on scientific findings.

METHODS

Protein Preparation. For comparative purposes, we used the same set of 10 kinase and 3 nuclear receptor targets as in our previous work.[12] The pdb codes and the names of the considered targets are as follows: 1ke5 (cyclin dependent kinase, CDK2), 1of1 (thymidine kinase), 1opl (c-ABL tyrosine kinase), 1p62 (deoxycytidine kinase, DCK), 1unl (cyclin dependent kinase, CDK5), 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.[22] 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), and then a receptor-grid was generated using default parameters.

Ligand Preparation. 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,[23] and then minimized in MOE with the MMFF94x[24−26] force field to a gradient of 0.0001 kcal/mol Å^2. We will refer to this ligand conformation (which was often significantly different from the X-ray structure) as the *seed conformation*. We showed in our previous work that the ligprep procedure, commonly used before Glide docking, does not produce superior results to this procedure and hence it was not applied in this work. The seed conformation was then used to generate ensembles for docking. The *dynamics ensemble* contained 50 conformations and was obtained after a short molecular dynamics simulation (300 K for 49.5 ps with sampling every 0.5 ps using a step size of 0.001 ps), followed by minimization of the sampled structures to a gradient of 0.1 kcal/mol Å^2 using the MMFF94x force field in MOE. Since the aim was only to generate alternate starting conformations for docking, these simulations were performed using distance-dependent solvation terms and default settings for the potential. The *conformational ensemble* was generated using the MOE Conformational Import application, starting from the seed conformation, using force field minimization with MMFF94x and a root-mean-square gradient termination criterion of 0.1 kcal/mol Å^2. The lowest energy conformers (maximum 50) within a 5 kcal/mol energy window were kept, hydrogens were added to these structures using the "Wash" procedure in MOE, and the structures were further optimized to 10^{-4} kcal/mol Å^2. Three further ensembles were employed in this work. Different *torsional grid ensembles* were created by generating conformations with uniformly spaced torsional angles around the seed conformation. The number of steps along in each direction was the same and was selected such that the total number of conformations was greater than 50 and as close to this number as possible. Subsequently, a diverse subset selection was performed to separate out the 50 most diverse conformations from this set. The *10°, 1°, and 0.1° torsional grid ensembles* covered a maximum sweep of 10°, 1°, and 0.1°, respectively, in the torsional search. Thus, the 0.1° torsional ensemble represents very small variations around the seed conformation; the corresponding structures are so similar that, when displayed together, they cannot be distinguished on a regular computer screen. To illustrate the conformational spread in these ensembles, Figure 2 shows the overlaid conformations for one of the targets (1ywr). Also note that the 1u4d ligand in our previous work has no freely rotatable bonds, and since these ensembles could not be generated for that ligand, it was not considered in this work.

Docking. The docking methods and settings used in this study were similar to those described previously.[12] Since our aim was to study the sensitivity of docking results to the initial input, most of the settings used were default. In the docking programs applied (GlideXP,[27] FlexX,[28] Surflex,[29] and GOLD[30]), we used the latest versions available to us. Since our previous report there have been two major changes in Glide that we thought could affect the results of this work: an optional ring conformational search was introduced that is applied during docking and an additional postdocking molecular mechanics minimization step was added to improve the complementarity of the ligand with the site. The effects of these changes on variability and score were also investigated. The linux version of Glide was used for almost all the experiments described here. Experiments were run on different SunFire and HP systems with 32- and 64-bit architecture under RedHat Enterprise 4−5.1, all producing identical results. The only exception is one experiment where both the linux and the Windows versions of Glide were used to dock the conformational ensemble of input structures to the VEGFR-2 target, in order to compare the variation in the top-scoring solution as produced by the different platform versions of Glide. The Windows runs were performed on a Dell Precision T7500 running under Windows XP.

Interpretation of the Results. Since the purpose of this work was to reduce the variability in the best score (and hence also the corresponding top-scoring pose) that arises from predocking input differences, the available scoring functions were accepted "as is", with no attempts at improvement or modification. Furthermore, it was assumed (as is with most commercially available scoring functions) that the "best" scoring pose is the "best" docked solution. Since we had no absolute measure as to what the best score should be for a given target, we accepted the top scoring solution found from all runs as the best score. Due to weaknesses in existing scoring functions, the best solution is not necessarily the most experimentally accurate, but the conclusions obtained here should hold, since further improvements in scoring functions will only mean increasing concordance between the best docked solution and experimental data. Since our aim is to understand the behavior of docking with respect to input variability, the issue of whether or not the highest scoring solutions were indeed the most accurate is beside the point.

RESULTS AND DISCUSSION

We have shown previously[12] that using different ligand conformations as docking input can produce top-scoring solutions that cover a wide range of score values, corresponding to several orders of magnitude in predicted ligand binding affinity. Given that the input ligand conformations are all valid starting points for docking, this was a startling outcome. This is especially so since docking validation studies are usually performed with only one ligand conformation as input[8] and, in practice, users typically expect the ideal docking scenario depicted in Figure 1a, namely, similar docking results for all valid inputs of the same molecule. On the basis of the limiting cases discussed above, if

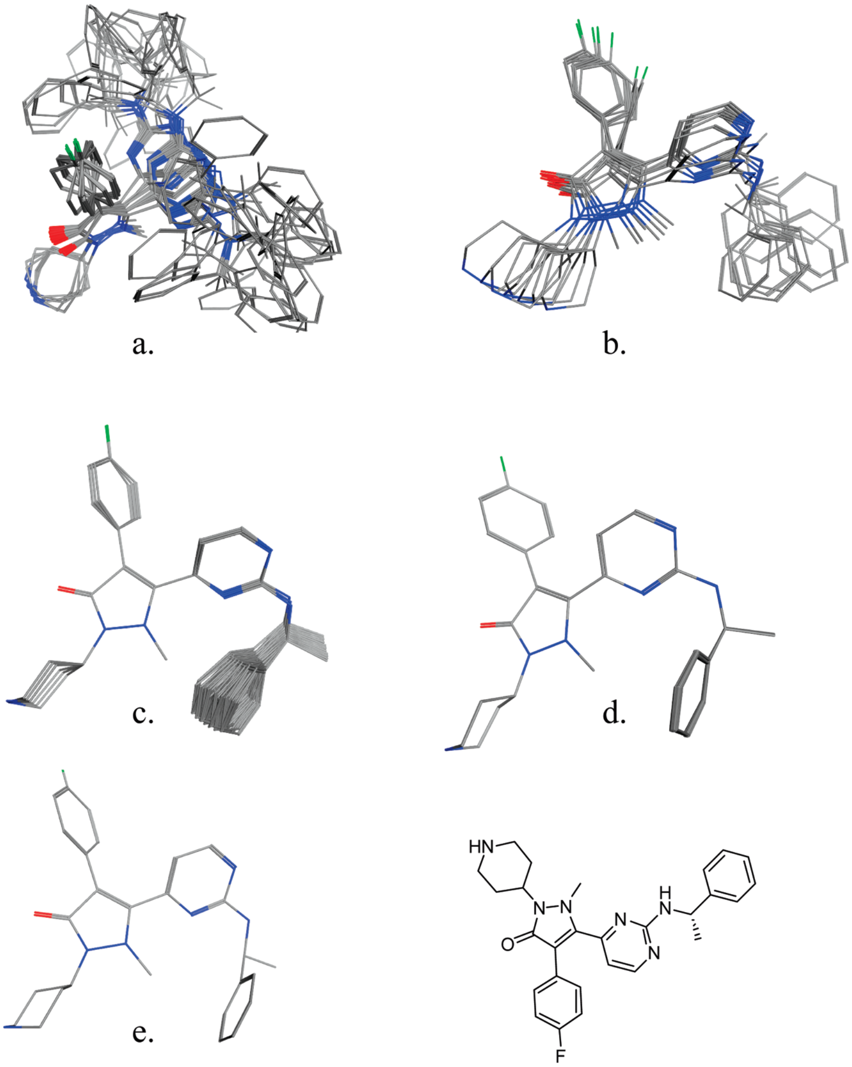

Figure 2. Overlaid ligand conformations used as starting points for docking to the 1ywr X-ray structure (p38 MAP kinase). All of these ensembles contain 50 conformers. (a) Conformational ensemble (CONF): lowest energy conformers, obtained after MM minimization following a conformational search. (b) Dynamics ensemble (DYN): structures obtained from a 49.5 ps MD run, followed by MM minimization. (c) 10° torsional grid ensemble: conformations obtained from a 10° angular sweep around the seed structure. (d) 1° torsional grid ensemble: conformations obtained from a 1° angular sweep around the seed structure. (e) 0.1° torsional grid ensemble: conformations obtained from a 0.1° angular sweep around the seed structure. The 2D structure of the ligand is shown for reference. See the text for further details about the generation of these ensembles.

incomplete ligand conformational searching by the docking engine is responsible for the variability in docking results, providing the docking engine with a diverse set of input conformations should complement the incomplete conformational search of the docking engine and produce better results. However, if the variability is caused by "chaotic effects", then the docking process becomes predominantly a "numbers game" and increasing the sheer number of input conformations (irrespective of conformational diversity) might improve the results.

We can gain some insight into the relative importance of these various effects by looking at the score ranges observed from all five ensembles considered, as shown in Figure 3. Most docking programs cover relatively larger score ranges when input geometries from conformational analysis are used, smaller ranges when these inputs come from dynamics, and even smaller ones when these inputs arise from an angular scan, with little discernible difference (apart from Surflex) between the 0.1°, 1°, and 10° scans. Such a behavior would indicate that score variations are more the result of inadequate conformational search (the search within the docking method is unable to make up for differences among the input conformers) and less caused by some random (chaotic) behavior. It is interesting to note that the lack of statistically significant difference between the three angular scans probably arises from the fact that even the 10° scan generates conformations that are "largely similar" (i.e., they are near the same energy minima) and the conformational search within the docking process has an easy task of bringing them back to the same docked conformations, with small differences arising mainly from chaotic effects. The only program with somewhat different behavior is Surflex, where there are large and "random" differences between the 0.1°, 1°, and 10° ensembles for different targets and the averages in Figure 3 give a relatively poor representation for the behavior of this program on any one target. We can also see in Figure 3 that both the absolute score ranges and their relative importance are different for different programs. Interestingly, the stochastic GOLD program consistently seems to produce the smallest score variations and the most

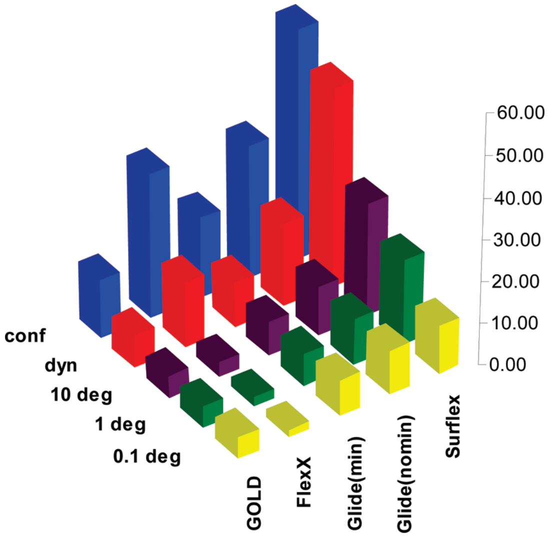

Figure 3. Relative score ranges, averaged over all considered targets. Score ranges in this graph were calculated for each method considering the same number of conformations, using the formula 100 × ($\text{score}_{\text{max}} - \text{score}_{\text{min}}$)/$\text{score}_{\text{max}}$ and then averaged over all targets. The large variations in the results from the conformational ensemble inputs seem to indicate that insufficient conformational coverage is the major factor for all docking programs (blue bars are larger than the other bars). However, as judged from the bars representing the 0.1°, 1°, and 10° ensembles, the stochastic/chaotic behavior is a significant factor for all of the programs, except for FlexX. The stochastic GOLD program is the only one that displayed consistently stable behavior across all ensembles. In the case of the Glide program, we can see a consistent improvement on applying ring conformer search and postdocking minimization [cf. Glide(min) vs Glide(nomin) columns].

robust output. From a development perspective, it is also interesting to compare the performance of Glide without postdocking minimization and ring search (default in 2007) and with these (default in 2009): the improvement in the consistency of results between versions is striking and is largely (∼80%) due to the ring conformational search they recently introduced but also to some extent the postdocking minimization (∼20%). These numbers are rough estimates only, obtained from the results when only one of the two effects was used in the calculations.

An interesting way to examine the relationship between the input ligand conformations and the final docking score and pose is to use time points from molecular dynamics simulations as inputs for docking. Conformations belonging to neighboring time points in the simulation are usually quite similar to each other, with occasional jumps to a different conformer. The plots in Figure 4 show the evolution of the best Glide docking score plotted against the corresponding simulation time point from which the ligand input geometry was obtained. The overlaid docked poses are also shown. The figure also includes the rms distance between the seed conformation and the MD ligand conformation at each time step *t*. This data is included to show the lack of correlation between the final score and the rmsd to the seed conformation. The results are given for three targets that exemplify the three types of behavior observed in this study. The first type of behavior, shown for the 1opk system in Figure 4a, was exhibited by the majority of targets considered. All top solutions in this case bind similarly to the target and the differences in the docking scores arise as a result of subtle differences in ligand conformation and pose. This seemingly random behavior with respect to docking score demonstrates the possible sensitivity of scoring function to minor variations in the docked geometry. This sensitivity is one source of score variation as a function of ligand input geometry.

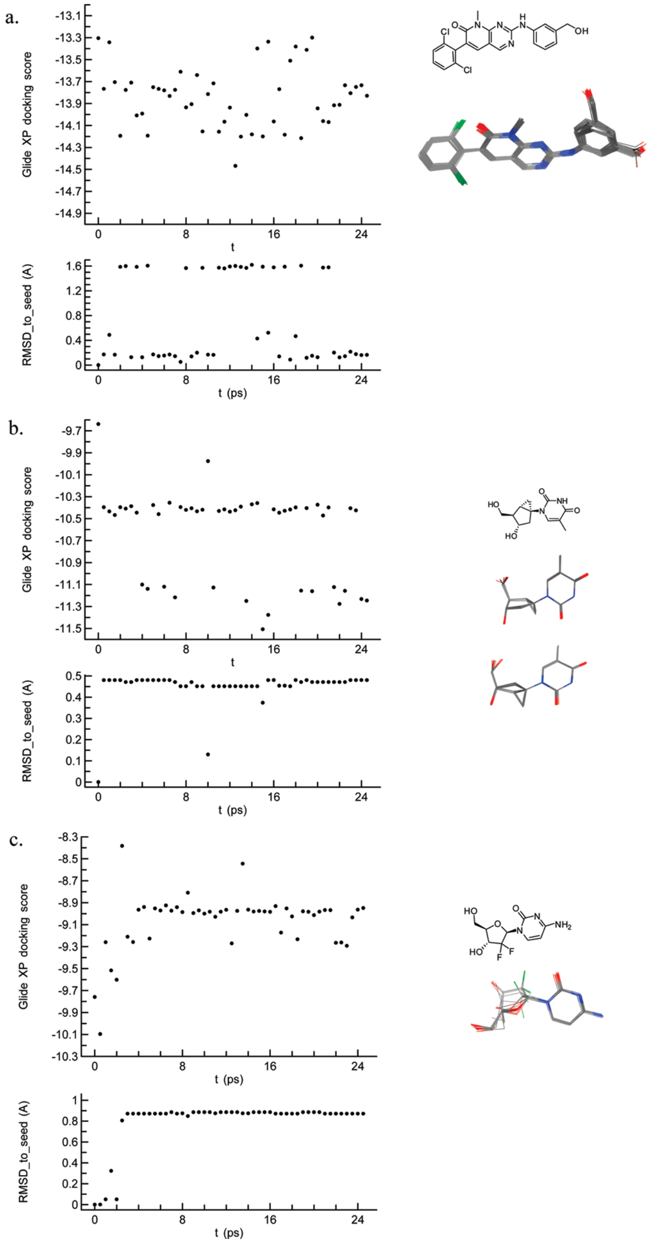

Figure 4. The relationship of simulation time in molecular dynamics and the GlideXP docking score exemplifying three major behaviors observed in this work. The ligand structure and the overlaid docked poses are shown beside each graph. The figure also includes the rms distance between the seed conformation and the MD ligand conformation at each time step indicating that this rms distance has little correlation with the final score. (a) Seemingly random behavior with scores obtained within a range of energies (1opk target). (b) Switching between two different scores, which correspond to two conformational states, arising from the flipping of the cyclopentane ring. The higher scoring set of solutions corresponds to the X-ray conformation (1of1 target). (c) Essentially one solution. The first few points from dynamics lead to docking solutions in good agreement with the X-ray conformation (the docked pose of the conformation from the 1.5 s time point is within 0.3 Å of the X-ray solution). Subsequently, the furanoid ring changes conformation in the dynamics simulation and docking is unable to bring it back, leading to a plateau in the score curve (1p62 target).

In a second type of behavior, depicted in the Figure 4b, two different binding modes with different scores arise from docking different input geometries. The plot shows random jumps between the two modes and again there are large score variations for the poses within each mode. (Note again that we are only dealing here with the top-scoring solutions.) If the underlying issue was conformational coverage alone, one would expect neighboring MD starting points to produce similar poses with similar scores. However, the results of this experiment show otherwise; similar input ligands can end up in different distinct docked solutions, suggesting that the pose search is incomplete and that the scoring function may be sensitive to small ligand perturbations.

The third type of observed behavior is shown in Figure 4c. This case (for the 1p62 target) is the only case where neighboring time points (i.e., similar conformations) produce more or less similar docking scores and poses. This may be due to the fact that the 1p62 binding site is a very tight pocket and docking the seed structure (which is the $t = 0$ time point in dynamics) correctly identified the binding conformation of the five-membered ring, seen from the cocrystallized ligand. In the course of dynamics, this ring conformation slowly changed and docking was unable to bring it back to the highest scoring ring conformation. This is another example of how incomplete conformational searching in the docking engine can affect results. From the examples studied here, it was the only one with more or less regular behavior, jointly arising from the tightness of the pocket and the inadequacy of the ring conformational search in docking.

The examples above demonstrate that similar input conformations can lead to highly dissimilar docked solutions. The scoring functions themselves can be quite sensitive to small positional variations (Figure 4a), leading to large score variations for similar poses, and similar input ligand conformers can indeed produce dissimilar poses. Thus, it appears from the above, at least for the Glide software, that sensitivity to initial conditions is indeed an effect in these docking runs. In the course of our experiments, similar behavior was observed using all of the described software, but to different extents. For example, FlexX tended to display behavior similar to Figure 4b,c, whereas Surflex was more likely to produce distributions similar to Figure 4a. GOLD results were generally also closest to Figure 4a but with much narrower score variations. As will be shown below, the explanation lies in the extent different docking software display the nonideal behavior shown by the limiting cases in Figure 1b,c.

To investigate the distribution of top scoring solutions over input conformation space, the input conformations were projected down to 2D with metric scaling, using the procedure described previously.[31,32] Metric scaling preserves the relative distance of the input conformations, so similar conformations are proximal in this plot. The plots in Figure 5 show 2D projections of the ligand input conformation space at different levels of magnification for the VEGF-R2 example. Each point represents a single ligand input conformation, with input conformations that produced docked solutions with scores in the best 20% (with Glide, using ring-search and minimization, scores < −13.7) shown in black. The first level of magnification shows the combined space of the conformational, 10° and 0.1° ensembles. The next level of magnification focuses on the 10° ensemble, while the final magnification level shows the 0.1° ensemble. Interestingly, the input ligand conformations that produce the highest scoring solutions are evenly dispersed throughout all three ensembles, as shown by the different magnification levels. This demonstrates that there are no clusters of similar input geometries that lead to better docking scores. Rather, the highest scoring solutions are obtained from diverse and dissimilar starting points, well-dispersed within the available conformational space. This suggests that increasing the number of starting conformations, at least to some extent, leads to improved results, because the number of starting points is higher, and not because the input conformations are more diverse. This conclusion is also supported by the fact that performing a 0.1°, 1°, or 10° conformational scan prior to docking led to no significant differences in the docking results, as long as the number of conformations was the same in these scans. The exact same behavior can be observed for other protein targets and ensembles. Again the

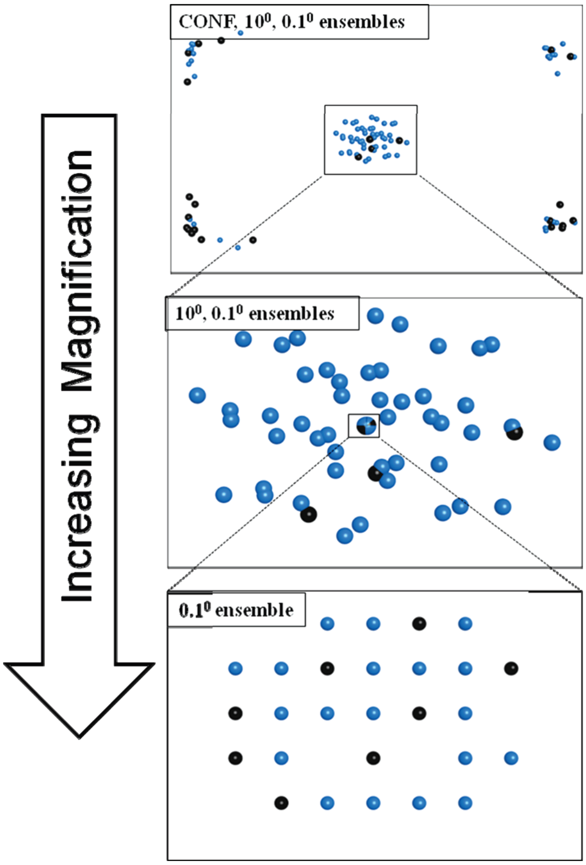

Figure 5. Distribution of the top docking scores in the input conformation space of the 1y6b ligand. The 3D conformations were projected down to 2D using metric scaling, preserving the relative distance of the conformations. The points are colored such that input conformations that produced the best 20% of docked scores (Glide with minimization and ring-search, score < −13.7) are in black, and other input conformations are in blue. The plot is shown at three levels of magnification. The top plot shows the combined conformational, 10°, and 0.1° ensembles. Here the 10° ensemble is confined to the center of the plot, while the 0.1° ensemble structures are so similar that they reduce to one point. The middle plot shows an intermediate magnification, focusing on the region of the 10° ensemble. In this plot, the 0.1° ensemble is still confined to a tiny region at the center of this plot. The final plot is at the highest level of magnification and shows a close-up of the 0.1° ensemble only. Note that the top scoring input structures (the black points) are distributed somewhat randomly over all three ensembles.

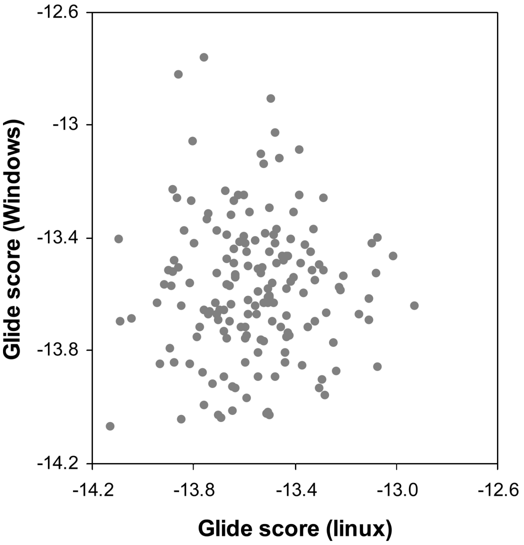

Figure 6. Correlation of the top scoring docked solutions produced by the linux and Windows versions of Glide, using identical docking settings and a diverse set of VEGF-R2 conformations as inputs. If there was no variation between the linux and Windows scores, a line of identity ($r^2 = 1$) would have been produced. However, the plot shows a great deal of variation in the top score (up to 1.0 unit of score) between the Windows and linux versions of Glide, even when the same input ligand conformation is used.

only exception is the 1p62 target; due to the limited conformational freedom, all solutions after metric scaling are located on a paraboloid and the high scoring solutions are all clustered on that line.

As a test of the variation in top scoring pose as produced by Glide under linux and Windows, each conformation from the conformational ensemble was docked with the linux and the Window version of Glide, and the top scoring solution was retained. A plot of the top score on linux versus the top score on Window is shown in Figure 6. If there was no variability between the program versions, the score vs score plot would be a line of identity; however, the plot in Figure 6 shows little correlation between the top scores produced by each program. The plot highlights the fact that a top score difference of up to 1 can be observed by running the same input on the linux vs the Windows version of Glide.

To further investigate the sensitivity of the deterministic docking programs to ligand input variation, we used the following automated procedure to get an average view of how docking results change with the number of different input conformations considered and the diversity of those conformations. From each ensemble of ligand input geometries (conformational, dynamics, 10°, 1°, and 0.1°), a random subset of a certain size (N_{conf}) was selected and the best scoring docked solution from that subset was found. This random selection was then repeated 500 times and the mean of the achieved best scores computed. This was repeated for different values of N_{conf}. The difference between the achieved best scores and the global-minimum best score ever recorded for that target−ligand pair [dS(max)] is then plotted against N_{conf}. Thus, these plots represent “on average” how far from the absolute best score one would be if N_{conf} different starting ligand geometries were used as docking input and the best score was taken from all of the N_{conf} docking runs. Although this procedure oversamples some of the smaller ensembles, the oversampling should not have any detrimental effect on the end result. In Figure 7, the mean deviation from the best score is plotted against the value of N_{conf} for the 1y6b target (VEGF-R2). Note that in all the Figure 7 plots, the deviation from the best score [dS(max)] decreases with increasing N_{conf}, meaning that increasing the number of input conformations leads to better scores. The best score obtained by a single run of the seed structure is also shown in the plots, denoted as a dotted horizontal line.

The plot in Figure 7a shows the results obtained using GlideXP with no ring conformational search and no post-docking minimization. Interestingly, for the torsion ensembles, dS(max) levels off between 0.3 and 0.5 at an N_{conf} value of ∼10. In contrast, the conformational and molecular dynamics ensembles continuously decrease and do not level off by $N_{\mathrm{conf}} = 20$. Furthermore, the 1° and 0.1° ensembles level off to different values of dS(max), which is surprising because these ensembles are essentially identical due to the small torsion angle variations. This plot suggests a docking scenario that is quite sensitive to initial conditions, as (i) similar ligands result in dissimilar docked scores and (ii) increasing the number of input structures increases the chances of finding the “global” minimum docked score. Introduction of ring searching and postdocking MM minimization to the GlideXP docking protocol improves the results significantly (Figure 7b). For all ensembles, using ring-search and minimization causes dS(max) to decrease more rapidly and even level off somewhat for the conformational and molecular dynamics ensembles. Furthermore, variations between the torsion ensembles almost disappear and the plots tighten up. These results suggest that the ring-search and minimization steps help to overcome some of the initial state sensitivity seen without these options. However, the ensembles do not converge to the same score, suggesting there is still an issue with completeness of the searching of pose space and score space.

The results using FlexX are somewhat different from the GlideXP results (Figure 7c). As with GlideXP, using more input conformations leads to better scores, but with FlexX, the leveling off of dS(max) happens more quickly than with GlideXP. The torsional ensembles all behave in a similar manner, suggesting little chaotic behavior (similar input conformations all produce similar docked poses), i.e., the score is not sensitive to small ligand pose perturbations. However, the conformational and molecular dynamics ensembles fail to find scores as good as those obtained with the torsion ensembles, suggesting that despite their greater diversity compared to the torsional ensembles, these ensembles do not contain a ligand geometry similar to those produced from the torsion ensembles and thus do not produce the same docked pose and minimum score. This behavior would pinpoint the coarse granularity of the conformational search as the culprit behind these variations.

The behavior of Surflex (Figure 7d) is yet a different scenario. The 0.1° and 1° torsion ensembles show almost identical behavior and level off to the maximum score (dS(max) = 0) quite quickly. This initially suggests a rather predictable (i.e., nonrandom) behavior, as similar ligands produce similar scores and poses. However, inspection of the 10°, conformational and molecular dynamics ensemble plots shows that none of these come close to converging to the best score, and none even seem to level off by $N_{\mathrm{conf}} = 20$.

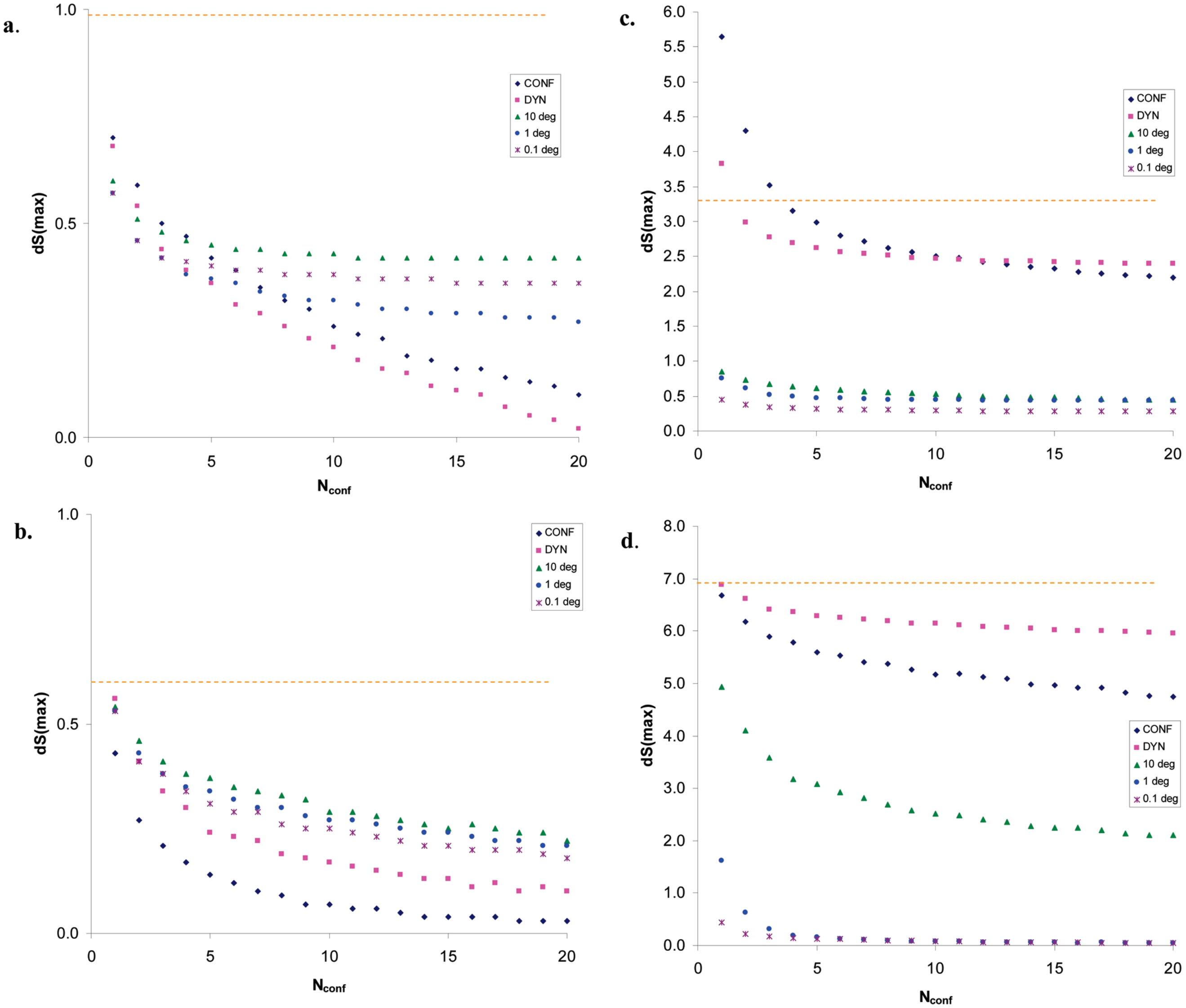

Figure 7. Dependence of the deviation from the best score seen for the target as a function of the number of solutions considered for the VEGF-R2 receptor. These plots were obtained by randomly selecting the given number of conformation (as shown on the *x*-axis) from the total number of conformations (maximum 20 considered) 500 times and calculating the mean. The orange dotted line represents the performance of the single MMFF-minimized Corina-generated (seed) conformation. (a) GlideXP with no ring conformational search during docking and no postdocking minimization (default settings for Glide 2007) (b) GlideXP with ring conformational search during docking and postdocking minimization (default settings for Glide 2009), (c) FlexX, and (d) Surflex.

It should be noted that the plots in Figure 7 are for a single target only and should not be used to make general statements about other targets and conformation generation methods. Generating plots similar to Figure 7 for all the targets shows notable differences between systems, namely, which conformer ensemble attains the best scores, how rapidly these curves reach the asymptote, and what the relative order is in terms of improvements in comparison to the seed conformation. Instead of presenting these plots separately for each target (which anyway would be misleading given the somewhat random nature of these results), it may be more instructive to plot the mean curves over the 13 targets studied, as shown in Figure 8. Note that since the number of conformations considered for different targets was slightly different, the score is normalized such that dS(max) for each target is divided by the number of conformations used for that target. Here, the value of dS(max)/N_{conf} vs N_{conf} is averaged over all targets in the study, and the results are plotted. In all cases the score improves with the number of input conformations N_{conf}, but there are some trends that should be noted. The GlideXP results in Figure 8a,b reflect the results from the single VEGF-R2 experiment; namely, inclusion of ring searching and postdocking minimization reduces the variation between ensembles and yields better scores with smaller values of N_{conf}. On the other hand, the results with the most recent Glide protocol (Figure 8b) show an interesting outcome, in that all three torsional and the dynamics ensembles converge similarly (i.e., the increase in the number of conformers from any of these ensembles have similar effects), but including a diverse set of conformers from the conformational ensemble provides a slight edge, pointing to the inadequacy of the conformational search as the major effect responsible for the variations.

The averaged FlexX results in Figure 8c closely mirror the FlexX results for the single VEGF-R2 target. The similar input conformations of the torsion ensembles converge to similar scores, suggesting a somewhat complete search of score space. However, the conformational and molecular dynamics ensembles consistently fail to produce the best docked score and level off at a best score that is about 2.5

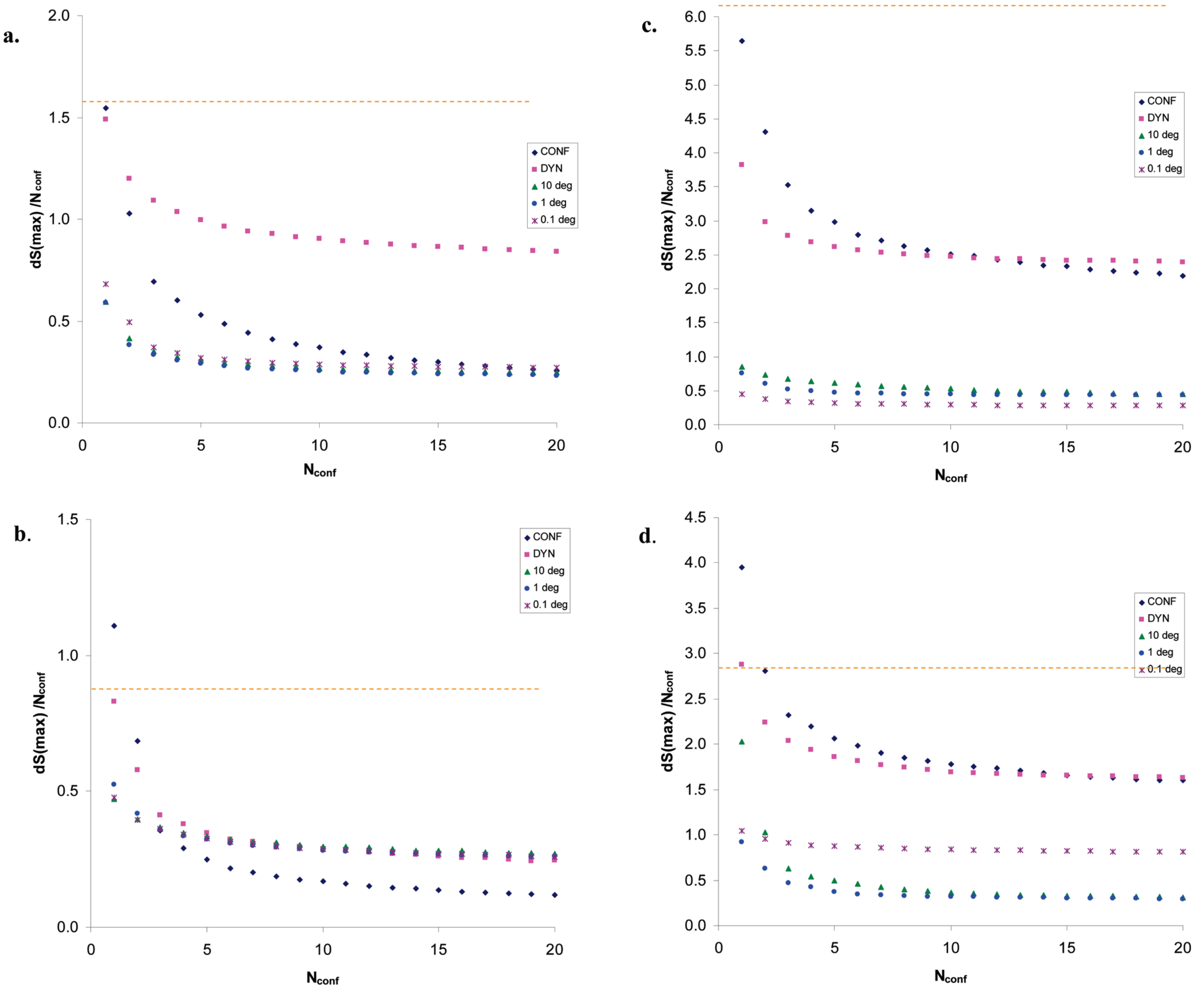

Figure 8. Mean deviation from the best score over different targets as a function of the number of solutions considered. These plots were obtained by randomly selecting the given number of conformation (as shown on the *x*-axis) from the total number of conformations (maximum 20 considered) 500 times, calculating the mean for each target, and then averaging over all the targets considered. The orange dotted line represents the performance of the single MMFF-minimized Corina-generated (seed) conformation. (a) GlideXP with no ring conformational search during docking and no postdocking minimization (default settings for Glide 2007), (b) GlideXP with ring conformational search during docking and postdocking minimization (default settings for Glide 2009), (c) FlexX, and (d) Surflex.

higher than the global minimum score. The averaged results for Surflex (Figure 8d) show a leveling off of the dS(max) after ~10 input conformations, but like FlexX, the conformational and molecular dynamics ensembles fail to find the minimum solution found by the 1° and 10° torsion ensembles. Interestingly, the 0.1° torsion ensemble also does not find the minimum solution, suggesting that there are various effects at play here that lead to score variability.

It may appear surprising that the position of the seed conformation is rather poor in all of these Figure 8 graphs; i.e., having more than one starting conformation seems to improve the results considerably in all cases. Another reason why the seed conformation is not among the best is the fact that all points in Figure 8 were obtained by calculating average errors for the displayed number of conformations, which provides more realistic results because averaging removes random effects. On the other hand, since the seed conformation is affected by random effects, it always fared worse than the averaged results for the different ensembles shown in these diagrams.

Using the data produced in the described docking experiments allows the breakdown of the docking score variability into two components. At first approximation we can assume that score variations produced from the conformational search ensembles speak to the inadequacy of the conformational search engine during docking, whereas variations produced from the nearly identical conformations of the 0.1° ensemble show the importance of “chaotic effects”. If we plot the variability (i.e., averaged relative ranges over all targets) along the *x*- and *y*-axes for the two sets, the obtained points for different docking engines will lie within a right triangle, as shown in Figure 9. In this case, fully ideal docking behavior (i.e., that corresponding to Figure 1a) with no variability in the docking score for either ensemble would be represented by the left-hand vertex of the triangle. If the variability from the 0.1° ensemble is identical to that of the conformational ensemble, the variability is mainly caused by stochastic effects (this would correspond to the hypotenuse of the triangle), whereas if the variability for the 0.1° ensemble is zero, the variability in the docking score would

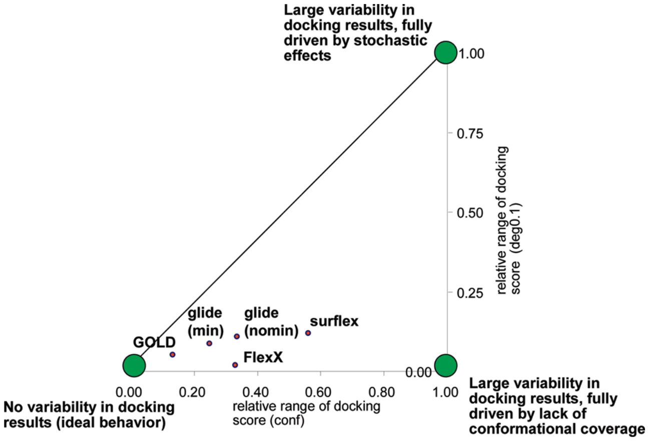

Figure 9. Breakdown of the docking score variability for different docking engines. The *x*-axis represents the variability for the conformational ensemble, the *y*-axis for the 0.1° ensemble. Fully ideal docking behavior would entail no variability in the docking score for either ensemble. If the variability from the 0.1° ensemble is identical to that of the conformational ensemble, the variability is mainly caused by stochastic effects (this corresponds to the hypotenuse of the triangle), whereas if the variability for the 0.1° ensemble is zero, the variability in the docking score is caused by incomplete conformational coverage (this corresponds to the *x*-axis side of the triangle). Using this representation and with the settings and targets applied in this work, the stochastic GOLD program appears to display the most robust behavior. Interestingly, FlexX appears to be the least influenced by chaotic effects, the variability with regard to input conformations in that program being mainly caused by incomplete conformational coverage.

primarily be caused by incomplete conformational coverage (this corresponds to the bottom side of the triangle). On the basis of this representation, we can compare the docking engines used in this work. The GOLD program appears to be closest to ideal behavior, which is in agreement with the findings of this and our previous study. This is rather surprising, given that GOLD is a stochastic program. Obviously, for this program alone, some of the variability would be caused by the genetic algorithm that would also manifest itself if identical inputs were used (See Table 1 in our previous study[12]). The diagram in Figure 9 also indicates that of the studied programs, FlexX appears to be the least affected by chaotic effects (possibly because of its incremental construction algorithm) and most of its variability is the result of the inadequacy of the conformational search. Note that FlexX has options to improve the granularity of this search, but these were not investigated in this work. Figure 9 also shows the effects of the improvements in Glide (introduction of ring conformational search and postdocking minimization) that helped to move this program closer to the ideal behavior.

On the basis of the results of this work, one could establish some practical guidelines (summarized in Table 1) that might improve the reproducibility and robustness of docking routines. Application of these guidelines is probably most beneficial when achieving the best accuracy of docking results is more important than the throughput of structures to be docked. (This would be the case during the lead optimization phase of drug discovery, when accurate ranking of potential ligand modifications is essential.) First, it is clear that the new Glide protocol with ring search and postdocking minimization is clearly beneficial in reducing the amount of variation in the docking results and should be utilized if possible. If these latter options are utilized, the shapes of the averaged curves in Figure 8b indicate that docking multiple structures from a limited conformational search prior to docking significantly improves the results. Note, however, that a single random conformation can be a much worse starting point than the minimized Corina structure or its equivalent, since a conformational search during docking might be unable to compensate for a highly different conformation. It would appear that introducing up to about five or six conformations has a dramatic effect on the quality of results and above ∼10 conformations we start to get diminishing returns. Thus, the optimum number of conformations might be somewhere around these numbers. If a predocking conformational search is not an option, we can still obtain significant improvement if we use multiple conformations, even if they are fairly similar (but they must not be identical). Ideally, a good starting point such as a minimized Corina or ligprep-generated structure should be included among these. The situation is very different for FlexX, where the torsional grid ensembles do not really help at all and using a diverse set of predocking conformations might produce the least amount of variability. Although it might appear from Figure 8c that the torsional ensembles produce better results than the conformational ensemble, the apparent difference is merely due to a single target (1sj0) where few of the inputs from the conformational ensemble lead to a correctly docked solution (the majority of these solutions are partly outside the pocket) whereas all of the inputs from the torsional ensembles do. Hence using three to five conformers from the conformational ensemble for FlexX appears to lead to sufficiently good results; higher numbers are not expected to be essential in most cases. It is more difficult to provide recommendations for Surflex: the strikingly different behavior of the 0.1° ensemble from the other torsional ensembles leads us to believe that the results might be too strongly affected by random variations. However, it is clear from the graphs that considering multiple input structures is expected to be beneficial in the Surflex calculations as well and probably a somewhat higher number of them would be necessary than with the other methods in this work. For GOLD, we did not find any advantage in using an ensemble of input conformations over simply increasing the number of dockings within the program. Note that the above recommendations were made on the basis of averages over all examples in this study, and within each example, the performance was averaged over many different possible selections. Thus, just to be safe, one might in practice use a somewhat higher number of conformations for best performance on specific targets. Since increasing the number of input conformations creates a high CPU burden for the docking process, it may be prudent to find a compromise between the number of ligand input conformers and the acceptable degree of variation on a per-target basis.

The rules-of-thumb presented in this paper are intended for use in lead optimization, where performing multiple docking runs on a series of similar compounds usually produces consistent poses and reasonable correlation between docking scores and relative activities within the series.[12] However, since docking scores across dissimilar chemical series typically do not correlate strongly with activity, these rules of thumb may not necessarily improve enrichment rates in high-throughput docking screens. Although performing

Table 1. Guidelines for Reducing Score Variations with Different Docking Programs Used in This Work[a]

	GOLD	Glide	FlexX	Surflex
minimum number of conformers to include	seed conformation sufficient	5−10	3−5	>10
source of conformations	not applicable	ideally from predocking conformational search; using other sources of conformation also beneficial	must be from predocking conformational search	any, but include some high quality starting points
other comments	increase number of dockings within the program to improve results, if necessary	use ring search and postdocking minimization options		

[a] Note that the guideline numbers above were obtained using averages over specific targets (kinases and nuclear receptors) and can be considered only as rough estimates. Also, they refer to the settings used in this work. These parameters also have major target-to-target variations and the effect of these is also different for different docking engines. Hence, if more consistent predictions are needed (e.g., in the lead optimization phase of drug discovery), it might be prudent to establish the optimal number of predocking conformations for the given target.

multiple runs will increase the chances of finding the best docked solution for each screened compound, the score differences between compounds from different chemical series may still not accurately reflect real activity differences due to the general inaccuracy of scoring functions, leading to compound misranking. As an aside, the research presented here has some profound implications for virtual screening protocols in general. Most large-scale virtual docking screens use only a single input conformation for each ligand. Depending on the protocol, the starting ligand conformation can be obtained from a wide variety of sources: SD files, SMILES converted to 3D, and 2D drawings converted to 3D using various programs and protocols. Since different ligand input conformations lead to different docked solutions, all virtual screens using docking are potentially sensitive to how the ligand input conformation was generated. Changing how the input ligands are produced could potentially have a large effect on the virtual screen results, although the effects of these changes are presently largely unpredictable.

CONCLUSIONS

In our previous work[12] we established that differences in the input conformation lead to variations in both the pose and the score of docking solutions. This dependence on input geometry is somewhat disturbing and should be reduced for various reasons, such as achieving better reproducibility and accuracy. In this work, we have tried to obtain a better understanding of the effects involved in order to find ways toward improvement. We broke down that variability from input variations to two independent effects: the inadequacy of the conformational search during docking and random “chaotic” effects. Using some approximate measures, we managed to plot the performance of different docking packages with these two effects as axes. This approach might be helpful for gauging the performance of different docking packages, which should be another perspective in comparative docking studies. Such plots should also be useful to establish if certain measures are helpful to reduce such effects (as shown for the improvement of Glide between two releases).

The two causes leading to the variability of docking results require different solutions. If the root cause of variability is the inadequate nature of conformational sampling during docking, the best solution might be to provide the program with a diverse set of conformations, dock all of them, and simply select the top scoring one. (Although not discussed here, we have shown that a similar approach is also useful in finding the most reliable estimate for other, higher energy solutions.) If, however, the cause of variability lies in chaotic effects alone, the process becomes a numbers game: in this situation, increasing the number of input structures will lead to improvements irrespective of the quality and diversity of these solutions (as long as they are valid conformations and not exactly identical). As was concluded from Figure 9, conformational sampling is usually the larger of the two effects. Thus, it is generally wiser to provide a small number of diverse conformations to the docking engine, rather than many similar ones. To that effect, guidelines for a reasonable number of input conformations were estimated for the docking programs studied here. However, target-to-target variation of these results cannot be overemphasized; hence, the prudent approach is to establish the optimum number of predocking conformations for the target of interest. The number of predocking conformations should also be considered in view of the required throughput of the docking engine. In particular, although the use of multiple input conformations should always be helpful, it is probably not warranted in lead-finding efforts and high-throughput docking exercises. In contrast, it is highly useful when the objective is to obtain quantitative rankings or identifying the highest scoring pose, as is often the case during the lead optimization stage of drug discovery. We have previously shown that this practice leads to improvements in the correlation of docking score and binding affinity. Indeed, it is hoped that such practices could, in general, contribute to the better use of docking in lead optimization.

ACKNOWLEDGMENT

Paul Labute of the Chemical Computing Group is gratefully acknowledged for his insightful suggestions. James Pan and Fuqiang Ban of the Campbell Family Institute for Breast Cancer Research are also acknowledged for helpful discussions.

REFERENCES AND NOTES

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nature Rev. Drug Disc.* **2004**, *3*, 935–949.

(2) Seifert, M. H. J. Targeted Scoring Functions for Virtual Screening. *Drug Discovery Today* **2009**, *14*, 562–569.

(3) Leimkuhler, B.; Chipot, C.; Elber, R.; Laaksonen, A.; Mark, A.; Schlick, T.; Schütte, C.; Skeel, R. Free Energy Calculations in Biological Systems. How Useful Are They in Practice? *Comput. Sci. Eng.* **2006**, *49*, 185–211.

(4) Stjernschantz, E.; Marelius, J.; Medina, C.; Jacobsson, M.; Vermeulen, N. P. E.; Oostenbrink, C. Are Automated Molecular Dynamics Simulations and Binding Free Energy Calculations Realistic Tools in Lead Optimization? An Evaluation of the Linear Interaction Energy (LIE) Method. *J. Chem. Inf. Model.* **2006**, *46*, 1972–1983.

(5) Rao, C. B.; Subramanian, J.; Sharma, S. D. Managing Protein Flexibility in Docking and Its Applications. *Drug Discovery Today* **2009**, *14*, 394–400.

(6) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(7) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies of Protein−Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(9) Davis, A. M.; St-Gallay, S. A.; Kleywegt, G. J. Limitations and Lessons in the Use of X-ray Structural Information in Drug Design. *Drug Discovery Today* **2008**, *13*, 831–841.

(10) Søndergaard, C. R.; Garrett, A. E.; Carstensen, T.; Pollastri, G.; Nielsen, J. E. Structural Artifacts in Protein−Ligand X-ray Structures: Implications for the Development of Docking Scoring Functions. *J. Med. Chem.* **2009**, *52*, 5673–5684.

(11) Enyedyi, I. J.; Egan, W. J. Can We Use Docking and Scoring for Hit-to-Lead Optimization? *J.Comput.-Aided Mol. Des.* **2008**, *22*, 161–168.

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

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

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

(15) Sornette, D. Nature Debates: Earthquakes. 1999. http://www.nature.com/nature/debates/earthquake/, last retrieved on May 7, 2010.

(16) Groison, D. Est-Il Vrai Que les Ordinateurs Font des Erreurs de Calcul. *Sci. Vie* **2002**, *1022*, 130.

(17) 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.

(18) Biesiadecki, J. J.; Skeel, R. D. Dangers of Multiple Time Step Methods. *J. Comput. Phys.* **1993**, *109*, 318–328.

(19) 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.

(20) Sandu, A.; Schlick, T. Masking Resonance Artifacts in Force-Splitting Methods for Biomolecular Simulations by Extrapolative Langevin Dynamics. *J. Comput. Phys.* **1999**, *151*, 74–113.

(21) 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.

(22) 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.

(23) *Corina, Version 3.2*; Molecular Networks GmbH: Erlangen, Germany, 2006.

(24) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1996**, *17*, 490–519, 520−552, 553−586, 587−615, 616−641.

(25) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1999**, *20*, 720–729, 730−741.

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

(27) *Glide, Version 2009*; Schrodinger Inc.: Portland, OR, 2009.

(28) *FlexX, Release 3.1*; BioSolveIT GmbH: Sankt Augustin, Germany, 2009.

(29) *Surflex, Version 2.11*; BioPharmics LLC: San Mateo, CA, USA, 2007.

(30) *GOLD, Version 4.1*; Cambridge Crystallographic Database: Cambridge, U.K., 2009.

(31) 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.

(32) 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.

CI100204X