

## Effect of Input Differences on the Results of Docking Calculations

Miklos Feher<sup>\*,†</sup> and Christopher I. Williams<sup>‡</sup>

Campbell Family Institute for Breast Cancer Research, University Health Network, Toronto Medical Discovery Tower, 101 College Street, Suite 5-361, Toronto, Ontario M5G 1L7, Canada, and Chemical Computing Group, Suite 910, 1010 Sherbrooke Street West, Montreal, Quebec H3A 2R7, Canada

Received February 22, 2009

The sensitivity of docking calculations to the geometry of the input ligand was studied. It was found that even small changes in the ligand input conformation can lead to large differences in the geometries and scores of the resulting docked poses. The accuracy of docked poses produced from different ligand input structures—the X-ray structure, the minimized Corina structure, and structures generated from conformational searches and molecular dynamics ensembles—were also assessed. It was found that using the X-ray ligand conformation as docking input does not always produce the most accurate docked pose when compared with other sources of ligand input conformations. Furthermore, no one method of conformer generation is guaranteed to always produce the most accurate docking pose. The docking scores are also highly sensitive to the source of the input conformation, which might introduce some noise in compound ranking and in binding affinity predictions. It is concluded that for the purposes of reproducibility and optimal performance, the most prudent procedure is to use multiple input structures for docking. The implications of these results on docking validation studies are discussed.

### INTRODUCTION

Protein–ligand docking programs have been used for virtual screening and compound ranking since the early 1980s,<sup>1</sup> and over 60 different docking software packages are currently available to researchers.<sup>2</sup> Although there still exist significant deficiencies in docking applications,<sup>2–5</sup> they have been successfully applied to many drug design projects over the past decade.<sup>6–9</sup> The majority of docking programs require as input a three-dimensional structure of the biological target (typically a suitably prepared X-ray structure) and a single minimized 3D conformation of the ligand to be docked. The docking process typically consists of two components—(1) *pose generation*,<sup>4</sup> which generates ligand conformations, places them in the binding pocket (and possibly adjusts protein geometry to allowed for receptor flexibility) and (2) *scoring*,<sup>10,11</sup> which attempts to quantify the quality of a docked pose with an approximate binding free energy ( $\Delta G$ ) calculation or some other type of empirical scoring function.<sup>12</sup> Since pose generation is essentially the problem of sampling a huge configuration space,<sup>13</sup> errors arising from pose generation can in principle be ‘solved’ by exhaustive sampling, limited only by the time and computational resources allotted for the process. A more confounding issue in docking has perhaps been the problem with scoring function accuracy. Ideally, a scoring function should compute the protein–ligand binding free energy ( $\Delta G$ ), but current scoring functions are insufficiently accurate to do so reliably.<sup>14</sup> As a result, much effort has been devoted toward improving scoring functions<sup>15,16</sup> and the accuracy of binding free energy calculations.<sup>17,18</sup>

In recent years many validation studies (see in refs 19 and 20 for the most recent list of studies) have been published comparing the ability of different docking programs to correctly predict bound ligand geometries and to rank compounds by biological activity. The accuracy of a docked pose is usually gauged by the root-mean square distance (RMSD) between the docked pose and the X-ray structure. Success in activity ranking can be quantified using a number of statistical measures such as enrichment factors and ROC or enrichment curves. The failure of docking to accurately predict docked poses and to correctly rank compounds by activity is often attributed to deficiencies in either the placement methods and/or the scoring functions.

Since most docking programs perform their own conformational search to generate ligand conformers, ligands are usually supplied as a single 3D conformation. Programs such as GOLD,<sup>21</sup> MOE,<sup>22</sup> and Glide<sup>23</sup> alter torsion angles during pose generation, while programs such as FlexX<sup>24</sup> and Surflex<sup>25</sup> employ a ligand rebuilding process to generate ligand conformations and poses. Most programs require that the input ligand structure has optimal bond lengths and angles, as these quantities are not varied during pose generation, and may only be modified during an optional molecular mechanics refinement stage. Interestingly, the user manuals of these programs do not specify that the input ligand needs to be in a low-energy conformation—being at a local minimum is all that is required. Thus, it is often assumed that docking results are somewhat independent of the input ligand conformation, provided the input conformation is a reasonable one. The GOLD user manual even states that “The ligand conformation will be varied by GOLD during docking. The starting conformation therefore does not matter”.<sup>26</sup>

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

<sup>†</sup> Campbell Family Institute for Breast Cancer Research.

<sup>‡</sup> Chemical Computing Group.

In practice, there are a number of ways to generate a 3D ligand conformation for input into a docking program; it can be built from scratch in a modeling package, read in as a 3D structure from a variety of common file formats (SD, PDB, TriposMOL2), read in as a SMILES string, and converted to a 3D structure on-the-fly or minimized with one of a number of possible force fields. Small bond length and bond angle variations, along with larger torsional variations, are expected between structures built using different methods. These variations in ligand input can potentially produce variations in the docking output, but there is no mention of this possibility in any of the software user manuals. To date, there are no systematic literature studies addressing these variations and their possible magnitude. Until recently, this issue was only mentioned in relation to Glide and only in obscure places,<sup>27,28</sup> acknowledging the fact that differences in input torsion angles might lead to small differences in the docking results, provided the molecule is a weak binder, the pocket is tight, or the ligand is very flexible. As this work demonstrates, large differences may be produced even if these criteria are not met. It must be noted that in a recent special issue of JCAMD on the evaluation of computational methods, a few papers commented on differences between different ligand preparation methods and indicated that these might have an effect on docking performance.<sup>20,29,30</sup> The recommendation was to standardize the input ligand structures<sup>20</sup> e.g. by using minimized Corina structures. It will be shown below that although this approach might indeed make docking results more reproducible, it is still unable to level the playing field when comparing different targets and docking programs. Also, such standardization might not be the best strategy for obtaining the highest quality binding mode or score predictions.

Stochastic docking programs employ random numbers in pose generation and thus produce different poses and scores from repeated runs of the *same* input ligand conformation. These dispersions in docking scores can be interpreted as uncertainties and should ideally be reported as error bars in validation studies. Although a certain degree of variation is expected from these stochastic programs, one would hope that repeatedly running the same input would produce roughly the same list of docked poses with a small dispersion in the docking scores. However, a recent study<sup>31</sup> of stochastic docking programs showed that repeated runs of the same input can produce substantial dispersion in the final poses and scores. Although accuracy can be improved by extracting top poses from multiple repeat runs (suggesting an issue with incomplete sampling) the authors conclude that dispersions in the docking scores can be large enough to affect the accuracy of the results and the scientific conclusions drawn from them.

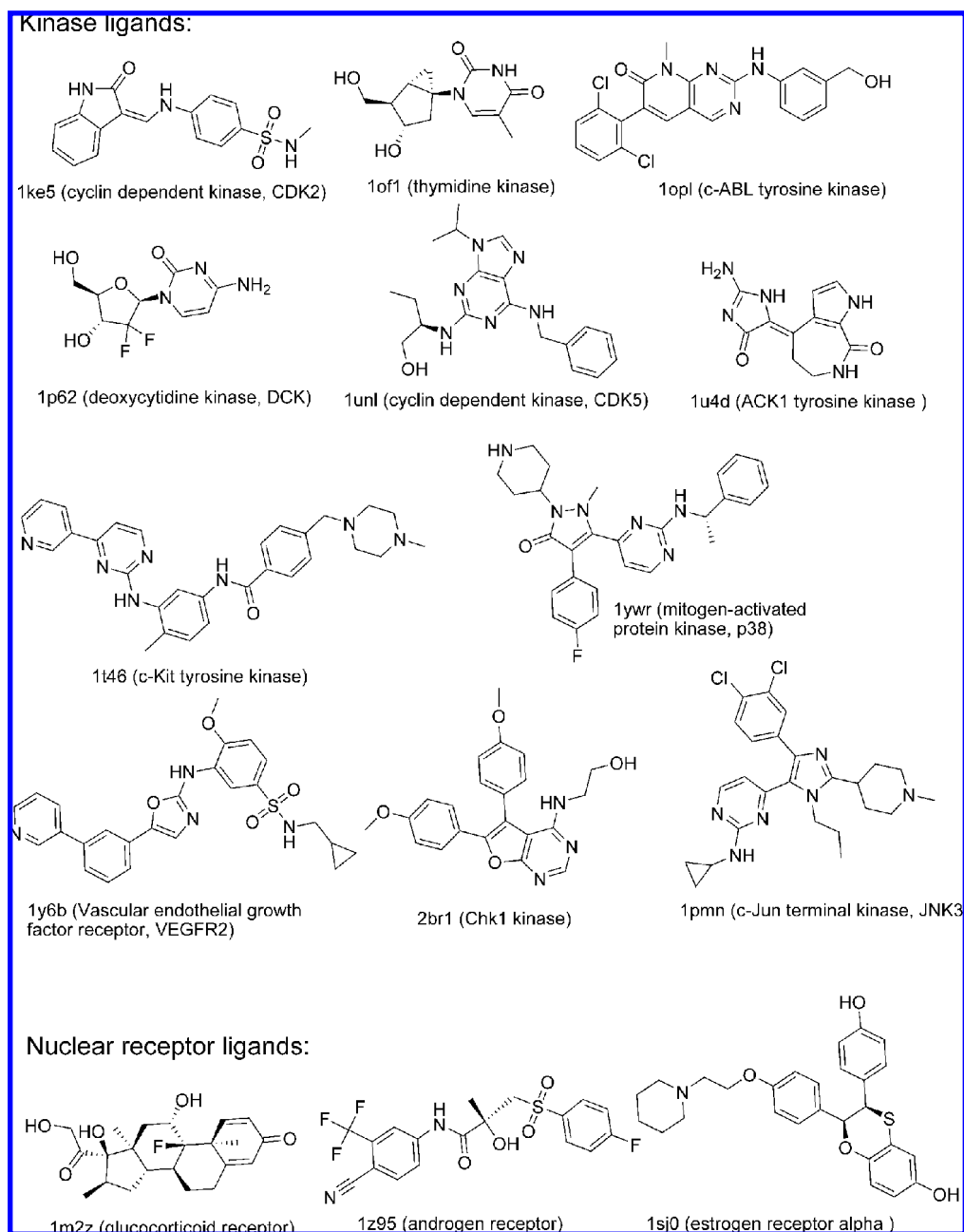
In contrast to stochastic programs, exact reproducibility is expected from *deterministic* docking programs which employ no stochastic elements in their searches—repeated runs of the same input on the same computer should produce identical output. Reproducible behavior of this sort is true of nonstochastic molecular mechanics minimizations, where repeated runs of the same input structure on a single-processor computer produce identical minimization trajectories leading to the exact same energy minimum.<sup>32</sup> However, it is known that even with deterministic algorithms, small input perturbations can generate large variations in the

final results. In a recent paper<sup>32</sup> we noted that very small changes in atomic input coordinates ( $<0.001$  Å) can have drastic effects on molecular mechanics minimization trajectories, resulting in large energetic and geometric differences between the final structures. In addition, different minimization results can be obtained if different computer hardware is used. It was concluded that with large systems such as proteins, molecular mechanics minimizations can exhibit *sensitivity to initial conditions*,<sup>33</sup> a state where each iterative cycle of an algorithm accumulates and magnifies small numerical variations to a point where the magnitude of the variation is comparable to that of the quantity being computed. Akin to minimizations, docking protocols are highly complex calculations with the potential of showing sensitivity to initial conditions. If molecular mechanics minimization is employed as a refinement step in the docking protocol, it is conceivable that the program already exhibits initial condition sensitivity due to the minimization step alone. The possibility of numerical instabilities in docking scoring functions has already been suggested by Aqvist et al.,<sup>34</sup> when they investigated the sensitivity of scoring functions to perturbations in the docked pose. In particular, they found empirical hydrogen-bond terms in the scoring function to be quite sensitive, with ligand coordinates deviations of as little as 0.5 Å or less leading to 3–4 kcal/mol differences in the estimated binding energy. Furthermore, differences in the treatment of random number sequences can give rise to variations in docking results obtained using different computer platforms. Within one computer platform/OS combination, a constant seed for the random number generator can be used to ensure reproducibility between repeated runs. However, the same seed may produce a different (but still reproducible) series of random numbers on a different computer platform/OS combination. In docking programs, conformation generation routines that use random numbers can thus be affected by random number generation differences between computer platforms (even when a constant seed is used), leading to variation in docking results between the computer platforms.

To examine the dispersion in docking results as a function of variations in the ligand input conformation, a representative set of commercial docking programs was chosen—GOLD, MOE, Glide, FlexX, and Surflex. A number of protein–ligand complexes were selected, and for each complex, different ligand conformations were used as input to docking runs, and the dispersion in resulting scores and docked poses were measured. Despite what may be expected, variations in 3D ligand input conformation can have drastic effects on the output list of docked poses and scores, a result that has profound implications for any virtual screening protocol that employs docking.

## METHODS

**Target Structure Preparation.** The targets selected for this work fell into two categories: kinases and nuclear receptors. We selected all of our targets from a curated set of 85 target and ligand structures that have recently been compiled and made publicly available.<sup>35</sup> Previous workers<sup>36</sup> have checked these protein structures for complete and unambiguous electron density maps, bad van der Waals clashes, and missing residues. All waters molecules were



**Figure 1.** Ligand structures (with target names and PDB codes) used in this study.

deleted from these structures. Despite this preparation, many of the structures still contained features (such as lone pairs) that could potentially cause errors in the software programs, so these were corrected in this work.

The chemical structures of the ligands from the selected complexes are displayed in Figure 1. Docking in this work was performed with GOLD 3.2, MOE 2008.10, Glide 4.5, FlexX 2.3, and Surflex 2.11. Most structure preparations and database manipulations were performed within the MOE software suite. Protein structures were first repaired and then properly protonated in the presence of its ligand using the Protonate3D<sup>37</sup> process in MOE. Proteins prepared in this manner were applied directly in FlexX, MOE, and Surflex. To be used in Glide, protein structures were further optimized using the protein preparation wizard in Glide (hydrogen-bond optimization and restrained structure minimization with the OPLS-AA force field to a maximum rms distance of 0.3 Å), and then a receptor-grid was generated using default

parameters. No protein preparation was performed for GOLD docking as the mol2 files specially prepared for this purpose were already published.<sup>36</sup>

**Ligand Input Conformations.** In normal drug discovery work the X-ray conformation of the ligand is unknown, and a 3D structure needs to be generated for docking. Since in this work we started from an X-ray structure, we applied a process to help ‘forget’ the X-ray ligand conformation—ligands were read into MOE, protonated/deprotonated using the Wash process, rebuilt into 3D using Corina,<sup>38</sup> and then minimized in MOE with the MMFF94x<sup>39–41</sup> force field to a gradient of 0.0001 kcal/mol Å<sup>2</sup>. This process also moves the ligand at least partially outside the site. The ligand conformation thus produced—henceforth referred to as the “seed conformation”—often differed significantly from the X-ray structure as a result of the rebuild process. Each seed conformation was used to generate an ensemble of conformations by performing a brief molecular dynamics simulation

(300 K for 49.5 ps with sampling every 0.5 ps and a step-size of 0.001 ps), followed by minimization of the sampled structures to a gradient of 0.1 kcal/mol Å<sup>2</sup> using the MMFF94x force field in MOE. Note that this is a somewhat greater than usual gradient value: it was chosen to ensure that differences between the starting structures were not completely eliminated. Thus for each target-ligand pair, an ensemble of 50 ligand input conformations was created—the initial seed conformation and 49 other conformations produced by minimizing structures from the dynamics runs. The seed conformation and the 49 other conformations together are henceforth referred to as the *dynamics ensembles*, which were used in the studies testing sensitivity of docking to input conformation.

In addition to the above preparation procedure, ligands to be docked with Glide were further prepared using the *LigPrep*<sup>23</sup> process (all options turned off other than the minimization using the OPLS-AA force field), and the outputs from docking were compared with and without LigPrep. In general, there was little difference between these results in terms of rms distances from the X-ray structures, although the results without LigPrep generally led to slightly lower conformational spread, more accurate geometries, and more consistent scores. However, as LigPrep is part of the usual ligand preparation for Glide, this method was applied in the majority of this work. The only data presented in this work without LigPrep is in Table 6. This was done for consistency with other methods and in particular to preserve the diversity of ring conformations from dynamics and conformational analysis.

**Ligand Conformational Analysis Ensembles.** Ligand ensembles were also generated using the MOE Conformational Import application, starting from the seed conformation. Force field minimization using MMFF94x was enabled, and a root-mean-square gradient termination criterion of 0.1 kcal/mol Å<sup>2</sup> was used. All conformers within a 5 kcal/mol energy window of the lowest energy conformation were kept. Next hydrogens were added to these structures (using the ‘Wash’ procedure in MOE), and the structures were further optimized to 10<sup>−4</sup> kcal/mol Å<sup>2</sup>. Ligand structures prepared in this manner are henceforth referred to as the *conformational analysis ensembles*.

**Docking Run Parameters.** Since the main goal of this study was to examine variations in docking results and *not* the accuracy, program settings that minimize the variation in results when repeatedly docking the same input file were chosen. Since these settings may differ from those used to optimize docking accuracy, the results presented here may differ from those reported in validation studies that seek to address docking accuracy. For the Surflex and FlexX programs, the default settings produced zero variation between repeat runs, so these were used in this study. Default settings were also applied in the GOLD program. These three programs were run on a HPxw8200 workstation with two Intel Pentium 4 CPUs at 3.2 GHz with 3Gb RAM running under Windows XP Service Pack 2. In the MOE software, the default nonstochastic Triangle Matcher placement method, followed by molecular mechanics refinement and GBVI scoring,<sup>42</sup> was used for the docking runs. The MOE runs were performed on a 1.6 GHz Sun SPARC computer with a sparcv9 dual-core processor, running the Solaris 10 operating system. For the Glide program, the extra

**Table 1.** Ligand Structure Preparation: RMSD Comparison of Ligand Seed Conformations with the X-ray Conformations and the Conformational Spread of the Dynamic and Conformational Search Ensembles<sup>a</sup>

PDB code and target <sup>b</sup>	X-ray vs seed conformation RMSD (Å)	dynamics ensemble pairwise RMSD: average (min, max) (Å)	conformation search ensemble pairwise RMSD average (min, max) (Å)
1u4d_ack1	1.12	0.11 (0.00, 0.001)	0.99 (0.42, 1.44)
1ke5_cdk2	0.89	0.13 (0.00, 0.66)	1.61 (0.31, 1.48)
1of1_thym	0.80	0.21 (0.00, 0.63)	1.24 (0.97, 1.54)
1p62_dck	1.58	0.33 (0.00, 1.13)	1.44 (0.26, 1.99)
1m2z_gr	0.42	0.38 (0.01, 0.72)	1.6 (0.36, 2.32)
1opk_abl	1.55	0.47 (0.01, 1.27)	1.42 (0.18, 2.26)
2br1_chk1	1.07	0.62 (0.01, 1.23)	1.16 (0.36, 1.65)
1y6b_vegf2	3.44	0.85 (0.01, 1.18)	2.75 (0.21, 4.69)
1pmn_jnk3	1.16	0.92 (0.01, 1.69)	1.9 (0.29, 3.05)
1z95_ar	2.45	1.12 (0.02, 2.45)	2.24 (0.64, 3.65)
1unl_cdk5	1.23	1.27 (0.26, 2.23)	1.75 (0.30, 3.04)
1ywr_p38	1.10	1.3 (0.01, 2.66)	2.43 (0.51, 3.88)
1sj0_er	2.05	1.71 (0.01, 2.56)	1.82 (0.41, 3.23)
1t46_ckit	1.86	1.89 (0.01, 2.31)	2.49 (0.30, 4.53)
<b>mean</b>	<b>1.48</b>	<b>0.81 (0.03, 1.48)</b>	<b>1.77 (0.39, 2.77)</b>

<sup>a</sup> Ligand seed is the conformation generated using Corina and subsequent force field minimization, whereas the dynamics ensemble was generated from the ligand seed using a short (49.5 ps) molecular dynamics simulation with 0.5 ps sampling. Values shown are averages, with the minimum and maximum RMSD distances given in brackets to illustrate the range. Further details are given in the text. <sup>b</sup> For the full name of these targets and a list of cocrystallized ligands, see Figure 1.

precision setting was enabled because this is the recommended setting for accurate docking. The Glide runs were performed on a Sun Fire V40z server with 8 dual-core ADM Opteron processors running under RedHat Enterprise Linux, release 4.

**Sensitivity to Input Conformations.** Each conformation from the dynamics ensembles was used as input into the docking programs. The highest scoring pose was kept from each docking list, resulting in 50 docked poses and scores for each target—one docked pose and score for each starting ligand conformation. The conformational spread in the docked poses was measured using average pairwise RMSD, while the score variation— $\Delta\text{Score}$ —is expressed as a ratio of the score standard deviation over one plus the absolute value of the average score (to avoid division by zero)

$$\Delta\text{Score} = 100 * (<\text{Score}>_{\text{STD}} / (1 + |<\text{Score}>_{\text{Average}}|))$$

The effect of input conformations on docking accuracy was tested using four different sources of 3D ligand input—the X-ray structure itself, the seed conformation, the entire dynamics ensemble, and the entire conformational analysis ensemble. In each case the highest scoring solution was kept, and its RMSD to the X-ray structure measured. From the dynamics and conformational analysis ensembles, the best scoring pose produced from the entire manifold of input conformations was chosen for RMSD comparison with the X-ray structure.

## RESULTS AND DISCUSSION

Each seed conformation was compared to the X-ray conformation from which it was generated by measuring the RMSD between the two optimally superposed structures. The resulting RMSDs in Table 1 show that some seed conforma-



**Table 2.** Absolute Reproducibility Tests: Average Pair-Wise RMSD of Ligand Poses and Score Variations Resulting from Repeated Docking Runs of the Same Input<sup>a</sup>

PDB code and target <sup>b</sup>	starting RMSD (Å)	Surflex		MOE		FlexX	
		pairwise RMSD (Å)	$\Delta$ Score (%)	pairwise RMSD(Å)	$\Delta$ Score (%)	pairwise RMSD (Å)	$\Delta$ Score (%)
1u4d_ack1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1ke5_cdk2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1of1_thym	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1p62_dck	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1m2z_gr	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1opk_abl	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2br1_chk1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1y6b_vegf2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1pmn_jnk3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1z95_ar	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1unl_cdk5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1ywr_p38	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1sj0_er	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1t46_ckit	0.000	0.000	0.000	0.000	0.000	0.000	0.000

PDB code and target <sup>b</sup>	starting RMSD (Å)	GOLD		Glide <sup>c</sup>	
		pairwise RMSD (Å)	$\Delta$ Score (%)	pairwise RMSD (Å)	$\Delta$ Score (%)
1u4d_ack1	0.000	0.080	0.200	0.000	0.000
1ke5_cdk2	0.000	0.070	0.310	0.020	0.260
1of1_thym	0.000	0.070	0.410	0.000	0.000
1p62_dck	0.000	0.090	0.470	0.000	0.000
1m2z_gr	0.000	0.080	0.120	0.000	0.000
1opk_abl	0.000	0.240	0.240	0.000	0.000
2br1_chk1	0.000	0.410	1.300	0.000	0.000
1y6b_vegf2	0.000	0.300	2.250	0.060	0.330
1pmn_jnk3	0.000	0.270	2.730	0.000	0.000
1z95_ar	0.000	0.410	1.360	0.070	0.000
1unl_cdk5	0.000	0.490	2.030	0.000	0.000
1ywr_p38	0.000	0.490	1.890	0.000	0.000
1sj0_er	0.000	0.270	1.600	0.000	0.000
1t46_ckit	0.000	0.280	0.990	0.000	0.000

<sup>a</sup> 10 identical copies of the input were docked in one batch. Ligand spreads were obtained as the average pairwise RMSDs of the final ligands, representing conformational and pose differences. Score variations are given by  $\Delta$ Score =  $100 * (<Score>_{STD} / (1 + |<Score>_{AVERAGE}|))$ . Further details are given in the text. <sup>b</sup> For the full name of these targets and a list of cocrystallized ligands, see Figure 1. <sup>c</sup> These Glide numbers were obtained on a single processor. For the three sulfonamides, only the first solution was different; the other 49 were identical. For consistency with the other methods, the numbers given are averages over the 50 structures.

tions are quite close to the original X-ray conformation (RMSD of 0.42–0.89 Å), while other seed conformations differ significantly from the X-ray conformation (RMSD > 2 Å). The conformational diversity of the dynamics and conformational ensembles can be assessed with the average, maximum, and minimum pairwise RMSDs, also listed in Table 1. Due to the short simulation time, in many cases the dynamics run produced an ensemble of very similar conformations (e.g., 1U4D, 1KE5, and 1OF1), with average pairwise RMSDs ranging from 0.11–0.21 Å, and maximum pairwise RMSD differences less than 1.0 Å. The conformational differences in these examples arise mainly from freely rotatable end-groups such as CH<sub>3</sub> and NH<sub>2</sub>. In some cases, the dynamics ensembles consisted of very dissimilar conformations, with large average pairwise RMSDs of up to 1.89 Å and maximum pairwise RMSDs > 2 Å. For all targets structures except 1UNL, the minimum pairwise RMSD of the dynamics ensemble is <0.03 Å, indicating at least one pair of near-identical conformers in these ensembles. In all cases the dynamics ensemble structures are minimized 3D conformations and hence satisfying the input criteria for various docking programs. As expected, the conformational analysis ensembles show wider conformational diversity than the dynamics ensembles. Compared with the dynamics ensembles, the conformational analysis ensembles exhibit larger maximum and average pairwise RMSDs values, while

the minimum pairwise RMSDs are almost always greater than 0.2, reflecting the fact that the application is intended to produce a diverse set of nonduplicate conformations.

To get an idea of the baseline variation in docking output, each program was run ten times using the only seed conformation as input, with the top scoring pose kept from each run. In Table 2 the results for ten repeated docking runs of the seed conformation are given. Since exactly the same input file was used for each run, the RMSD between the 10 starting ligand conformations is always 0.00 Å. The average pairwise RMSD between the top pose produced by each docking run of the same program are also listed. The results clearly show that the Surflex, FlexX, and MOE programs exhibit completely reproducible behavior across all targets; repeated runs of the same input file will produce exactly the same output of docked ligand poses with zero variation in the RMSD and zero variation in the docked ligand scores. Table 2 results also clearly show the stochastic nature of the GOLD docking program, as repeated runs of the same input in each case produces modest pairwise RMSDs and nonzero variations in docking scores between the top docked poses.

The results for the Glide program are not as clear as the other four programs; it seems to behave reproducibly for most targets, with 0 Å RMSD between the top docked pose from each run and no score variation, but on occasion Glide would

produce different results from the same input. For the 1KE5, 1Y6B, and 1Z95 targets (where the ligands contain a sulfonamide group), Glide shows some small variations in the output. Interestingly, these variations were only present in the very first solution if run as one batch on a single processor or in all the first solutions on each processor in a distributed *paraglide* run. Without stochastic elements, the production of different results from repeated runs of Glide with exactly the same input can only be explained by irreproducibility in some low-level operation or as a the result of a programming glitch.

Computer architecture, math libraries, and compiler options can all potentially introduce variations into the docking results. The completely reproducible MOE results reported in Table 2 were obtained on a Sun Sparc computer using an executable compiled with the Sun Studio 11 compiler. Similar reproducibility can be observed when running the same experiment on a Mac computer. However, nonzero RMSD and score variations are observed when the same experiments were performed on Intel chips with executables complied with an Intel compiler (in both Windows and Linux operating systems). While most of the variations were small ( $<0.001$ ), in one case (1y6b) an RMSD variation of  $>2$  Å was observed, as two distinct poses (corresponding to two separate potential energy wells) were produced from the replicate runs. Further investigation revealed that the observed variations with the Intel executable arise from the refinement step, as the docking placement step shows zero variation on all computer systems. Since the same underlying molecular mechanics MOE C code is used in all systems, differences in behavior between the executables can only be attributed to platform, compiler, and math library differences. Pinpointing the exact source of the Intel variations—the compiler, the math libraries, and/or the chip itself—is beyond the scope of this work. This result merely demonstrates that identical C code can potentially produce different results depending on the compiler used, the math libraries chosen, and the processor used for the execution.

In Table 3 the average, maximum, and minimum pairwise RMSDs of the dynamics ensemble are compared before and after docking. The table is sorted in the same order as Table 1—by increasing average pairwise RMSD of the dynamics ensemble before docking. Given that these docking programs perform their own ligand conformational searches and are meant to produce the best bound ligand pose from any reasonable 3D ligand structure, one might expect only small RMSD differences in the top poses produced from each input conformation. Indeed, in almost all cases, a minimum pairwise RMSD of  $<0.2$  Å between docked poses was produced, indicating that some instances different starting conformations can lead to near-identical docked solutions. However, in many cases the average pairwise RMSD of the ligand ensemble after docking (Table 3) is at least as large—if not larger—than the average pairwise RMSD before docking. The large average RMSDs of the dynamics ensembles after docking (Table 3) show the degree of variation between the top poses produced by different starting conformations. Furthermore, the maximum pairwise RMSDs of the dynamics ensembles after docking (Table 3) can be larger than 7 Å, showing that vastly different docked solutions can be produced from different input conformations. In only one instance—the FlexX program with the 1U4D system—did

**Table 3.** Average Pair-Wise RMSD (Å) within the Dynamics Ensemble before and after Docking with Various Commercial Packages, with Minimum and Maximum Pair-Wise RMSDs Shown in Brackets

PDB code and target <sup>a</sup>	dynamics ensemble (before docking)	Surflex	MOE
1u4d_ack1	0.11 (0.02, 0.25)	0.49 (0.00, 5.19)	1.55 (0.00, 2.01)
1ke5_cdk2	0.13 (0.02, 0.59)	2.53 (0.02, 8.48)	0.80 (0.00, 2.81)
1of1_thym	0.21 (0.04, 0.63)	0.31 (0.01, 1.07)	1.40 (0.00, 2.37)
1p62_dck	0.33 (0.04, 1.13)	0.53 (0.01, 2.40)	0.13 (0.00, 4.29)
1m2z_gr	0.38 (0.02, 0.84)	0.45 (0.01, 0.93)	0.18 (0.00, 0.69)
1opk_abl	0.47 (0.07, 1.27)	5.13 (0.02, 11.37)	2.59 (0.00, 3.47)
2br1_chk1	0.62 (0.04, 1.23)	5.50 (0.06, 9.06)	3.50 (0.00, 6.72)
1y6b_vegf2	0.85 (0.08, 1.45)	7.28 (0.06, 12.78)	2.87 (0.01, 7.89)
1pmn_jnk3	0.92 (0.08, 1.92)	5.07 (0.07, 8.58)	0.68 (0.01, 6.96)
1z95_ar	1.12 (0.07, 3.21)	1.86 (0.01, 4.92)	1.55 (0.00, 3.25)
1unl_cdk5	1.27 (0.26, 2.29)	3.86 (0.42, 8.49)	2.00 (0.01, 6.12)
1ywr_p38	1.30 (0.03, 2.73)	4.24 (0.02, 8.79)	3.90 (0.00, 7.11)
1sj0_er	1.71 (0.06, 3.83)	5.98 (0.05, 11.25)	0.77 (0.01, 4.17)
1t46_ckit	1.89 (0.06, 3.28)	5.93 (0.03, 16.83)	0.78 (0.00, 1.56)
<b>mean</b>	<b>0.81 (0.06, 1.76)</b>	<b>3.51 (0.06, 7.87)</b>	<b>1.62 (0.00, 4.24)</b>

PDB code and target <sup>a</sup>	FlexX	GOLD	Glide <sup>b</sup>
1u4d_ack1	0.00 (0.00, 0.00)	0.17 (0.02, 0.45)	0.97 (0.00, 3.61)
1ke5_cdk2	0.28 (0.0, 1.87)	0.17 (0.03, 1.08)	1.07 (0.03, 2.03)
1of1_thym	0.66 (0.00, 7.58)	0.31 (0.02, 3.85)	0.20 (0.01, 0.86)
1p62_dck	1.63 (0.00, 9.99)	0.30 (0.03, 1.33)	0.43 (0.01, 1.47)
1m2z_gr	0.63 (0.00, 1.11)	0.16 (0.02, 0.46)	0.23 (0.01, 0.51)
1opk_abl	0.37 (0.00, 1.67)	0.94 (0.04, 1.74)	1.35 (0.02, 2.50)
2br1_chk1	3.83 (0.00, 8.63)	0.80 (0.07, 1.48)	1.47 (0.06, 7.55)
1y6b_vegf2	1.29 (0.00, 2.79)	0.55 (0.12, 1.17)	0.59 (0.09, 1.06)
1pmn_jnk3	1.42 (0.00, 3.25)	0.37 (0.06, 1.55)	0.95 (0.04, 1.99)
1z95_ar	0.59 (0.00, 1.71)	1.03 (0.13, 1.65)	0.84 (0.06, 1.34)
1unl_cdk5	1.69 (0.26, 8.84)	0.89 (1.20, 1.55)	1.07 (0.15, 2.14)
1ywr_p38	0.75 (0.00, 4.29)	3.62 (0.12, 9.06)	0.93 (0.08, 1.47)
1sj0_er	3.08 (0.00, 10.25)	1.03 (0.07, 1.83)	1.83 (0.02, 12.41)
1t46_ckit	6.07 (0.00, 13.52)	0.89 (0.08, 1.56)	0.80 (0.03, 1.29)
<b>mean</b>	<b>1.59 (0.02, 5.39)</b>	<b>0.80 (0.14, 2.05)</b>	<b>0.91 (0.04, 2.87)</b>

<sup>a</sup> For the full name of these targets and a list of cocrystallized ligands, see Figure 1. <sup>b</sup> Using the recommended LigPrep protocol before Glide docking.

the ensemble of docked ligand poses exhibit an average pairwise RMSD of 0.0. This is the only example in this study of a docking result that is *independent* of the starting conformation; i.e. the same top-scoring docked pose is produced from each of the different input conformations. In all the other cases studied here, the nonzero average and maximum pairwise RMSDs between top docked poses indicate that different top poses are generated from each 3D ligand input. These results demonstrate that docking results are indeed sensitive to the initial ligand conformation and that all of these programs can produce different lists of docked poses and scores depending on the input ligand conformation. One striking example of this is the Surflex result for 1KE5; despite a small variation in ligand input structures (average pairwise RMSD of 0.13 Å), the resulting docked poses have an average pairwise RMSD of 2.53 Å.

**Score Variation.** In addition to variations in docked geometries, different ligand input conformations produce substantial variations in the final docking scores. Direct comparison of docking scores is tricky (because each score has a different physical meaning), so the previously described score variation ( $\Delta Score$ ) was used to compare variations between methods. In Table 4 the variations in docked scores over the 50 runs from the dynamics ensembles are given. One important observation is that the scores vary greatly, much more than the variation in RMSD values may lead us to expect. Ideally, if a docking program is not affected by

**Table 4.** Score Variation in Docking Runs:  $\Delta\text{Score}^a$  Obtained from Docking the Dynamics Ensembles




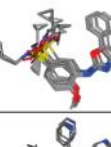
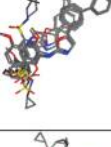
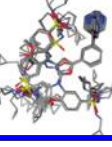
PDB code and target <sup>b</sup>	Surflex	MOE	FlexX	GOLD	Glide <sup>c</sup>
1u4d_ack1	7.55	7.12	0.01	0.18	1.01
1ke5_cdk2	10.84	6.16	1.85	0.40	1.71
1of1_thym	3.16	4.06	9.55	1.29	8.28
1p62_dck	8.17	8.54	5.36	2.00	2.55
1m2z_gr	4.42	0.66	9.20	1.28	0.94
1opk_abl	28.39	0.77	1.59	0.23	0.60
2br1_chk1	5.10	4.72	5.10	0.88	10.31
1y6b_vegf2	8.70	5.46	1.94	1.36	3.11
1pmn_jnk3	19.10	11.34	12.95	1.01	2.85
1z95_ar	19.64	0.61	4.55	2.15	1.33
1unl_cdk5	16.36	4.81	3.12	3.03	6.10
1ywr_p38	19.38	9.21	6.28	6.52	4.06
1sj0_er	16.03	9.72	3.10	4.43	15.29
1t46_ckit	21.47	12.56	5.05	1.36	3.09
<b>mean</b>	<b>13.45</b>	<b>6.12</b>	<b>4.98</b>	<b>1.86</b>	<b>4.37</b>

<sup>a</sup>  $\Delta\text{Score} = 100 * (\langle\text{Score}\rangle_{\text{STD}} / (1 + |\langle\text{Score}\rangle_{\text{AVERAGE}}|))$ . <sup>b</sup> For the full name of these targets and a list of cocrystallized ligands, see Figure 1. <sup>c</sup> Using the recommended LigPrep protocol before Glide docking.

the 3D ligand input conformation, different ligand starting conformations should lead to the same or very similar docking scores, and the resulting score variations will be near zero. In only one case—1U4D with FlexX—was there no variation in the score. For all other examples there are nonzero variations in the docking scores. The SD/mean ratio is in many cases larger than 10%. Note that 10% variation in the SD/mean ratio might hide differences as large as 50% in the absolute value of the score which, in the case of scores that are meant to correlate with binding energy, could mean several orders of magnitude difference in predicted  $K_i$ . Surprisingly, the smallest average score variation was observed for the GOLD program, which has stochastic elements, while the largest average score variation is observed for the deterministic Surflex program. These results demonstrate that the starting ligand geometry can have a profound effect on the final docking score.

To get a better feel for the variation in ligand geometries before and after the docking runs, the VEGF-R2 example will be discussed in some more detail. Table 5 compares an overlay of the 50 input structures (the dynamics ensemble) with overlays of the docked solutions. The RMSD and score variations also given. From the overlay of input structures (MD) we can see that in this case, dynamics has negligible effect on the overall structure, producing an ensemble of starting ligands with an average pairwise RMSD of 0.85 Å and a maximum RMSD of 1.45 Å. Visually, the 50 overlaid dynamics ensemble structures look quite similar, except for some variation in torsion angles of peripheral groups—specifically the pyridine rotation and the rotation of the ethyl chain that connects the cyclopropyl ring to the core. However all deterministic programs produce an ensemble of docked ligands with an average pairwise RMSD > 1.0 Å, i.e., greater than that of the input structures. Surprisingly, the only docking program that produced an RMSD spread less than that of the input ensemble was the GOLD program, which contains stochastic elements. The FlexX results explore a larger range of conformations for the end-groups, and substantial differences appear in the Glide results, as the entire end-groups move and are swapped in some solutions.

**Table 5.** Comparison of Different Docking Solutions for the VEGF-R2 Receptor: Average Pair-Wise RMSD and Score Variation of Dynamics Ensembles after Docking<sup>a</sup>

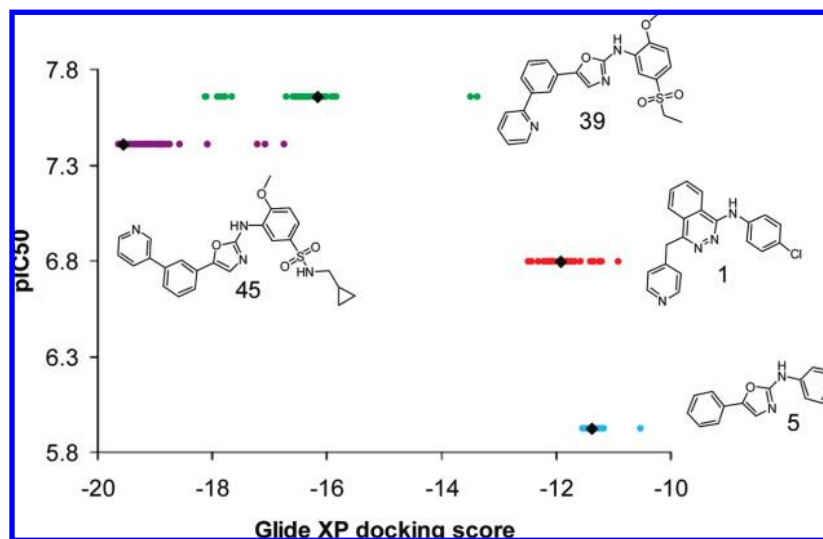
Method	Overlay of 50 solutions <sup>b</sup>	Ensemble RMSD (Å)	Score Variation (%)
Dynamics Ensemble (input)		0.85	-
GOLD		0.55	1.4%
GlideXP		2.17	3.1%
FlexX		1.29	1.9%
MOE		2.87	5.5%
Surflex		7.28	8.7%

<sup>a</sup> The dynamics ensemble (the initial seed conformation and 49 copies from a 49.5 ps molecular dynamics simulation) were docked to the VEGF-R2 receptor (pdb code 1Y6B). Score variation was calculated as  $\Delta\text{Score} = 100 * (\langle\text{Score}\rangle_{\text{STD}} / (1 + |\langle\text{Score}\rangle_{\text{AVERAGE}}|))$ . Further details are provided in the text. <sup>b</sup> For docking results the poses and conformations of the 50 solutions are overlaid as seen in the binding pocket. In the case of the dynamics results, these solutions were rigidly overlaid to minimize the rms distance of the corresponding atoms. For Surflex results, a similar rms-based alignment was performed for the picture above in order to clarify the nature of the obtained conformations.

MOE docking explores a large range of conformations with only a few torsion angles being left intact. The most striking example is the 7.28 Å average pairwise RMSD produced by the Surflex program; Surflex behaves deterministically when there is zero variation in ligand input geometries but appears to be very sensitive to variations in the ligand input geometry, and little similarity can be seen among any of the final solutions produced by Surflex. Given that these wildly different solutions are produced from a small set of starting points, these are rather striking results. This shows that relatively small changes in the starting geometry may lead to much larger changes in the final geometry after docking.

In order to demonstrate the effect of score variations on activity predictions, four inhibitors of the VEGF-R2 receptor were considered.<sup>43</sup> These ligands were treated similarly to other examples in this paper (3D structure generation by Corina, 50 copies of the input generated by a 24.5 ps molecular dynamics simulation followed by a subsequent

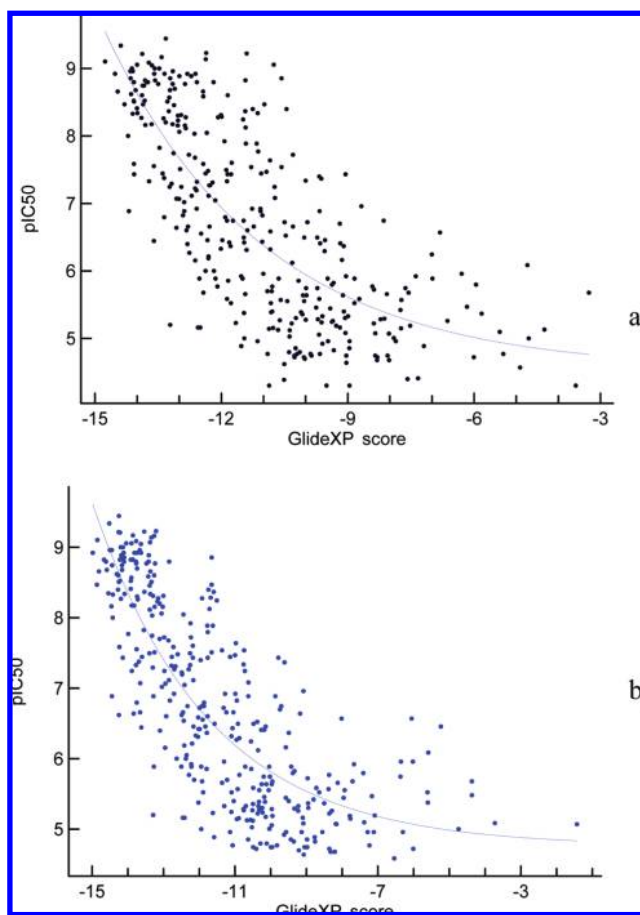




**Figure 2.** Glide XP docking scores vs experimental  $pIC_{50}$  for four VEGF-R2 ligands. As in other cases, 50 copies of the input were generated for each ligand using a 24.5 ps molecular dynamics simulation and docked using Glide XP. With the experimental  $pIC_{50}$  values on the vertical axis, the docking scores for each compound lie on a vertical line. Due to the dependence on input structure, these scores cover a large range of values, and these ranges often overlap with each other. Results obtained directly from the minimized Corina input structure (the *seed conformation*) are shown as black diamonds.

minimization). In Figure 2 the experimental  $pIC_{50}$  data are plotted on the vertical axis versus the Glide score on the horizontal axis. The Glide program was selected for this experiment because it is the program most commonly used for activity predictions in industry. The 50 calculated docking scores for each of the four compounds lie on four different horizontal lines. From the plot it is clear that the docking scores cover a fairly large range of values, and the ranges overlap. The score for the  $t = 0$  solution (i.e., the *seed conformation*) is plotted as a black diamond and can be anywhere between the minimum or maximum score. One extreme example of the spread in scores is compound **39**—despite being the most active compound, the large range of possible scores means that it could easily be ranked well below the less potent compound **45**. The lowest scores of **39** are similar to the highest scores of compound **1**, suggesting that in some cases compounds **1** and **39** may be given very similar rankings, despite the large activity difference between them. Furthermore, the large overlap in the score ranges demonstrates the possibility of reversing the predicted rankings if different starting structures are used. Depending on the input ligand geometry, it is conceivable that running a single docking calculation on each of the compounds **1** and **5** could produce the lowest observed score for compound **1** and the highest observed score for compound **5**—a result that would reverse the ranking of the compound relative to experiment. This result demonstrates that ranking ligands with docking can be quite arbitrary and depends heavily on the starting ligand conformation. Such behavior can present major problems for activity predictions and may be one of the factors responsible for the poor correlations with experimental data observed for many docking projects.

In order to illustrate this point, results are presented in Figure 3 for a proprietary protein kinase target using high quality  $IC_{50}$  data. The molecules cover a range of chemotypes arising from both screening different compound collections and targeted medicinal chemistry. In Figure 3a Glide docking was performed on minimized Corina structures (seed structures), which corresponds to an industry-standard protocol.



**Figure 3.** Glide XP docking scores vs experimental  $pIC_{50}$  for a proprietary protein kinase target. a) docking the minimized Corina structure (seed) and b) docking all low energy conformations (below 1 kcal/mol) after minimization and choosing the top scoring solution for each molecule. It appears from this and other similar examples that pregenerating conformations helps to reduce the number of compounds that fit the binding pocket but fail to dock and somewhat improves binding mode predictions and the general predictivity of scores.



**Table 6.** RMSD (in Å) of Different Docking Solutions from the X-ray Structure, Using Different Starting Points for Docking<sup>a</sup>

PDB code and target <sup>b</sup>	Surflex				MOE				FlexX			
	seed	conf	MD	X-ray	seed	conf	MD	X-ray	seed	conf	MD	X-ray
1u4d_ack1	1.12	0.25	1.12	<b>0.00</b>	0.68	0.68	0.71	<b>0.62</b>	1.11	0.25	1.11	<b>0.07</b>
1ke5_cdk2	<b>1.11</b>	1.20	1.20	1.20	<b>0.98</b>	1.25	1.66	1.24	0.40	0.38	1.64	<b>0.47</b>
1of1_thym	0.56	0.43	<b>0.18</b>	0.40	0.52	0.43	1.80	<b>0.42</b>	1.36	1.17	<b>0.48</b>	1.94
1p62_dck	0.23	1.31	0.23	<b>0.12</b>	0.65	1.08	0.98	0.78	<b>0.29</b>	1.87	0.38	0.88
1m2z_gr	0.59	<b>0.40</b>	0.77	0.71	0.53	1.61	0.46	<b>0.46</b>	0.45	0.52	0.30	<b>0.45</b>
1opk_abl	1.67	<b>0.45</b>	0.68	0.75	<b>0.31</b>	0.51	0.51	0.52	0.56	<b>0.43</b>	1.66	1.73
2br1_chk1	1.06	<b>0.74</b>	1.06	1.28	1.23	<b>1.07</b>	1.94	1.21	0.81	1.08	1.07	<b>0.80</b>
1y6b_vegf2	1.09	2.95	<b>0.04</b>	0.37	<b>1.18</b>	1.23	1.64	1.29	1.14	<b>0.79</b>	1.21	3.28
1pmn_jnk3	<b>0.62</b>	0.93	0.70	1.31	<b>0.77</b>	1.66	1.58	1.06	0.89	1.58	<b>0.89</b>	0.85
1z95_ar	<b>0.31</b>	0.40	0.32	0.26	0.78	0.99	0.84	<b>0.53</b>	0.47	1.10	1.07	0.68
1unl_cdk5	0.72	0.63	<b>0.54</b>	0.70	1.43	1.53	1.83	<b>0.99</b>	0.74	0.94	0.98	0.93
1ywr_p38	<b>0.33</b>	1.78	0.64	0.45	<b>1.51</b>	1.67	1.54	1.86	0.74	0.61	0.78	<b>0.29</b>
1sj0_er	2.39	2.53	2.45	<b>0.34</b>	2.47	<b>1.33</b>	1.48	1.27	2.44	1.51	2.56	<b>0.65</b>
1t46_ckit	1.39	0.36	1.31	<b>0.25</b>	0.29	1.43	0.39	0.43	2.24	<b>0.82</b>	0.86	1.76
<b>mean</b>	0.94	1.03	0.80	<b>0.58</b>	0.95	1.18	1.24	<b>0.91</b>	0.97	<b>0.93</b>	1.07	1.05

PDB code and target <sup>b</sup>	GOLD				Glide <sup>c</sup>			
	seed	conf	MD	X-ray	seed	conf	MD	X-ray
1u4d_ack1	1.12	0.25	1.12	<b>0.00</b>	1.12	1.12	1.12	<b>0.00</b>
1ke5_cdk2	0.27	0.45	0.27	<b>0.17</b>	0.34	1.23	1.23	0.41
1of1_thym	0.53	1.18	0.20	<b>0.09</b>	0.50	0.10	0.20	<b>0.01</b>
1p62_dck	<b>0.23</b>	0.84	0.87	0.30	0.26	1.31	0.26	<b>0.25</b>
1m2z_gr	0.21	0.48	<b>0.20</b>	0.27	0.56	0.63	<b>0.50</b>	0.53
1opk_abl	<b>1.64</b>	1.67	1.68	1.71	1.65	1.65	1.66	<b>0.50</b>
2br1_chk1	<b>0.84</b>	0.91	0.98	0.98	0.69	<b>0.63</b>	0.69	1.13
1y6b_vegf2	0.52	<b>0.25</b>	0.28	0.41	0.69	0.64	0.55	<b>0.19</b>
1pmn_jnk3	<b>1.39</b>	1.58	1.41	1.56	1.74	1.58	1.74	<b>0.66</b>
1z95_ar	0.53	0.79	0.79	<b>0.39</b>	0.58	0.34	<b>0.24</b>	0.26
1unl_cdk5	0.44	0.70	0.55	<b>0.24</b>	0.72	0.61	0.60	<b>0.47</b>
1ywr_p38	0.90	<b>0.56</b>	0.62	0.72	0.71	0.47	0.60	<b>0.16</b>
1sj0_er	2.38	1.07	2.46	<b>0.70</b>	2.55	1.48	2.47	<b>0.37</b>
1t46_ckit	0.55	0.44	0.58	<b>0.41</b>	0.29	0.71	0.75	0.29
<b>mean</b>	0.82	0.80	0.86	<b>0.57</b>	0.89	0.89	0.90	<b>0.37</b>

<sup>a</sup> Docking was performed with starting conformations from the minimized Corina structure (seed), the experimental X-ray conformation (X-ray), the ensemble of low energy conformers (conf), and from the ensemble of conformers obtained from a 49.5 ps molecular dynamics simulation of the seed structure (MD). In the latter two cases, the solution with the highest score was selected. See further details from the text. Best RMSD to the experimental X-ray conformation are in **bold**. <sup>b</sup> For the full name of these targets and a list of cocrystallized ligands, see Figure 1. <sup>c</sup> In order to preserve differences between input conformations and for consistency with other methods, the LigPrep protocol was not used before Glide docking.

A linear model describing the correlation has an  $r^2$  of 0.50 and a root-mean square error of 1.05 log units for the 330 molecules that docked. (The fit appears to be nonlinear, fitting e.g. a simple exponential decay curve results in an  $r^2$  of 0.57 and a root-mean square error of 0.99 log units.) To obtain Figure 3b docking was preceded by a conformational search, keeping all conformers within a 1 kcal/mol window (see the full description of the process above for the conformational ensemble). These structures were all docked, and the highest scoring docked solution was considered. This process provided some modest improvement statistically: a linear fit produces an  $r^2$  of 0.55 and a root-mean square error of 0.94 log units for the 352 molecules that docked. (Again, with the same simple exponential decay curve as above, an  $r^2$  of 0.66 and a root-mean square error of 0.82 log units is obtained.) More importantly, however, 22 more molecules docked with the prior conformational search that failed to do so without it. The binding mode of the docked structures is also generally more consistent with the expected hinge-binding if a prior conformational search was performed. We found similar behavior also for other enzyme targets. More systematic studies of this effect should lead to improved activity predictions. However, the main conclusion from

these results supports the major point of this work; namely, the nature of the input 3D ligand conformation can have a profound effect on docking results.

The results of this study have demonstrated that docking accuracy can be greatly influenced by the input ligand conformation. However, it is unclear which source of input ligand conformation leads to the most accurate docking results. To test the effect of the ligand input conformation on docking accuracy, the top-scoring pose produced by docking runs starting from one of the four different input sources—the seed conformation, the dynamics ensemble, the conformational analysis ensemble, and the X-ray structure—was retained and compared to the X-ray pose. This was performed for all the targets in the study. The RMSD to X-ray of the highest scoring pose from each run are listed in Table 6, with the RMSD value closest to experiment highlighted in **bold**. It must be emphasized here that the goal is not to determine which docking code is more accurate but rather to examine the effect of different conformations on the variation in docking results. The input parameters were not optimized for any of these programs, instead default parameters were applied unless there was an indication from the manuals to do otherwise (as in MOE and Glide). There were also biases built into protein preparation

for some methods but not for others: the protein files were originally carefully prepared for GOLD docking,<sup>36</sup> and specific protein preparation procedures were applied for Glide and MOE, which are expected to favor the results obtained by these methods.

If one considers only the mean RMSDs in Table 6, it appears that using the X-ray structure as the starting ligand conformation generally produces the docked poses closest to experiment. (It needs to be mentioned that the protein preparation wizard in Glide performs a limited minimization of the protein geometry together with the cocrystallized ligand, hence the X-ray structure has an *a priori* advantage as a starting point in Glide docking experiments and this bias also improves the apparent performance of Glide in comparison to other methods.<sup>44</sup>) However, the averaging above glosses over the fact that on a case-by-case basis, no one source of ligand conformation consistently produces the best pose for all targets. In fact, for all the docking programs in this study, every one of the ligand generation methods produces the best docked pose for at least one target, and there is a pretty even distribution for the sources of the best pose. Often it is assumed that starting with the X-ray conformation will give the best docking results, but inspection of Table 6 shows that in many instances using the X-ray ligand can produce the worst docked pose. This is despite the fact that these X-ray structures were included in the training sets of most of the applied programs. These results suggest that no one method of generating starting ligand conformations can guarantee the production of the best possible docked pose. Indeed, the 'best' starting ligand conformation depends on the target and the docking program.

## CONCLUSIONS

In this paper, we have shown that contrary to what one may expect, the poses and scores produced by docking programs can be heavily dependent on 3D ligand input structure. Relatively small and seemingly unimportant changes to the ligand input conformation can have drastic effects on the resulting poses and scores, even in cases where there are no stochastic components in the docking program. This result has many implications for any docking-based virtual screening protocol or comparative docking study that uses only one input ligand starting conformation and performs only one docking run per ligand. It also has implications on current efforts to improve activity predictions from docking. From the results in Table 6, it is interesting to note that having a full coverage of conformations prior to docking does not necessarily improve docking accuracy. In addition, there seems to be no ligand starting geometry that guarantees the best docking pose will be produced. Docking with the X-ray conformation as input can produce worse results than docking a random conformation from a dynamics run. These results raise two questions about what may be reasonable to expect from docking predictions; first, it might be misleading to expect good predictions straight from a single structure and, even after generating multiple starting points it is unclear whether we select the best one based on score. Second, the score varies greatly even with very similar starting conformations. Indeed we observed in several of our drug discovery projects that the docking

scores of nanomolar inhibitors might be high in some docking runs and low in others and the only way we could minimize this variation and have better reproducibility was to use multiple ligand conformations as starting points for each ligand. In our experience, such reproducibility (as tested for hundreds of drug-size ATP competitor molecules acting on kinase targets) can be achieved either by using over 5 ligand conformations from a molecular dynamics simulation or a conformational search preceding the docking where all solutions below 1 kcal/mol are kept (see the Methods section for further parameters). These however are minimalistic suggestions and, as shown in this work, do not necessarily produce the best results, only results that are less affected by the ligand starting geometry. Thus it appears that the prudent strategy may be to always use multiple starting points especially if quantitative activity predictions are sought. Without them the docking results can be strongly dependent on the starting geometry of the ligand. However it is still unclear how the best solution from the resulting spread of docked poses could be selected. Clearly, some improvement in the quality of scoring functions would be required such that the highest scoring solution from the spread would be nearest to the experimental structure.

## ACKNOWLEDGMENT

Paul Labute of the Chemical Computing Group is gratefully acknowledged for his insightful suggestions and lively discussions about the nature of this problem and potential sources of these effects.

## REFERENCES AND NOTES

- (1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–88.
- (2) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.
- (3) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
- (4) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409–443.
- (5) Hassan, S.; Gracia, L.; Vasudevan, G.; Steinbach, P. J. Computer simulation of protein-ligand interactions: challenges and applications. *Methods Mol. Biol.* **2005**, *305*, 451–492.
- (6) Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. Docking: Successes and Challenges. *Curr. Pharm. Des.* **2005**, *11*, 323–333.
- (7) David, L.; Neilsen, P. A.; Hedstrom, M.; Norden, B. Scope and Limitation of Ligand Docking: Methods, Scoring Functions and Protein Targets. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 275–306.
- (8) Good, A. Structure-based virtual screening protocols. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 301–307.
- (9) Blake, J. F.; Laird, E. R. Pharmacophore-based virtual screens. *Annu. Rep. Med. Chem.* **2003**, *38*, 305–314.
- (10) Boehm, H.-J.; Stahl, M. The use of scoring function in drug discovery applications. *Rev. Comput. Chem.* **2003**, *18*, 41–87.
- (11) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
- (12) Jain, A. N. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407–420.
- (13) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.* **2002**, *24*, 1637–1656.
- (14) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today* **2004**, *1*, 231–239.

- (15) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.
- (16) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **2006**, *11*, 421–428.
- (17) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371*, 1118–1134.
- (18) Gilson, M. K.; Zhou, H.-K. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
- (19) Warren, G. L.; Peishoff, C. E.; Head, M. S. Algorithms and Scoring Functions; State-of-the-Art and Limitations, in Computational and Structural Approaches to Drug Discovery: Ligand-Protein Interactions. *R. Soc. Chem.* **2007**, 137–154.
- (20) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
- (21) *GOLD, version 3.2*; Cambridge Crystallographic Database: Cambridge, U.K., 2008.
- (22) *MOE Molecular Operating Environment, version 2008.10*; Chemical Computing Group Inc.: Montreal, Canada, 2008.
- (23) *Glide, version 2007*; Schrodinger Inc.: Portland, OR, USA, 2007.
- (24) *FlexX, version 2.3*; BioSolveIT GmbH: Sankt Augustin, Germany, 2007.
- (25) *Surflex, version 2.11*; BioPharmics LLC: San Mateo, CA, USA, 2007.
- (26) *GOLD User Guide & Tutorials*; Cambridge Crystallographic Database: Cambridge, U.K., 2005.
- (27) Glide Frequently asked questions, question 32, available on [www.schrodinger.com](http://www.schrodinger.com) (accessed November 4, 2008).
- (28) *Glide Technical Notes for version 3.5*; Schrodinger Press: Portland, OR, 2005.
- (29) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
- (30) Jain, A. N.; Nicholls, A. Recommendations for evaluations of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
- (31) Onodera, K.; Satou, K.; Hirota, H. Evaluations of Molecular Docking Programs for Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1609–1618.
- (32) 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.
- (33) Lorenz, E. *The essence of chaos*; University of Washington Press: Seattle, WA, USA, ISBN: 0295975148.
- (34) Marelius, J.; Ljungberg, K. B.; Aqvist, J. Sensitivity of an empirical affinity scoring function to changes in receptor-ligand complex conformations. *Eur. J. Pharm. Sci.* **2001**, *14*, 87–95.
- (35) 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.
- (36) Database was downloaded from the CCDC Web site. [http://www.ccdc.cam.ac.uk/products/life\\_sciences/GOLD/validation/downloads/download.php4](http://www.ccdc.cam.ac.uk/products/life_sciences/GOLD/validation/downloads/download.php4) (accessed November 20, 2007).
- (37) Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 187–205.
- (38) *Corina, version 3.2*; Molecular Networks GmbH: Erlangen, Germany, 2006.
- (39) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1996**, *17*, 490512, 520–552, 553–586, 587–615, 616–641.
- (40) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1999**, *20*, 720729, 730–741.
- (41) Unpublished modification to MMFF94s; enforces planarity of conjugated nitrogens.
- (42) Labute, P. The Generalized Born/Volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration using London Dispersion Instead of Atomic Surface Area. *J. Comput. Chem.* **2008**, *29*, 1693–1698.
- (43) Harris, P. A.; Cheung, M.; Hunter, R. N.; Brown, M. L.; Veal, J. M.; Nolte, R. T.; Wang, L.; Liu, W.; Crosby, R. N.; Johnson, J. H.; Epperly, A. H.; Kumar, R.; Luttrell, D. K.; Stafford, J. A. Discovery and evaluation of 2-anilino-5-aryloxazoles as a novel class of VEGFR2 kinase inhibitors. *J. Med. Chem.* **2005**, *48*, 1610–1619.
- (44) Jain, A. Bias, reporting and sharing: computational evaluation of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.

CI9000629