JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Assessing Molecular Docking Tools for Relative Biological Activity Prediction: A Case Study of Triazole HIV-1 NNRTIs

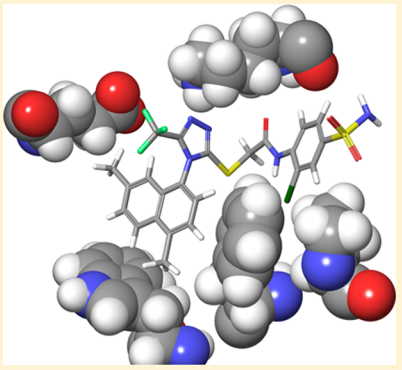
Tomasz Frączek,[†] Agata Siwek,[†,‡] and Piotr Paneth*[,†]

[†]Institute of Applied Radiation Chemistry, Lodz University of Technology, Zeromskiego 116, 90-924 Lodz, Poland

[‡]Department of Organic Chemistry, Faculty of Pharmacy, Medical University, Chodzki 4a, 20-093 Lublin, Poland

Supporting Information

ABSTRACT: Molecular docking is a technique widely used in drug design. Many studies exist regarding the general accuracy of various docking programs, but case studies for a given group of related compounds are rare. In order to facilitate identification of novel triazole HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs), several docking and scoring programs were evaluated for their ability to predict relative biological activity of 111 known 1,2,4-triazole and 76 other azole type NNRTIs. Glide, FlexX, Molegro Virtual Docker, AutoDock Vina, and Hyde were used. Different protocols, settings, scoring functions, and interaction terms were analyzed. We have found that the programs performance was dependent on the data set, indicating the importance of choosing good quality target data for any comparative study. The results suggest that after optimization and proper validation, some of the molecular docking programs can help in predicting relative biological activity of azole NNRTIs.



INTRODUCTION

With increasing computer speeds and the rising number of molecular modeling tools, along with a better understanding of ligand−receptor interactions, computer-aided drug design gains more and more significance. Today's medicinal chemists have a large selection of various computational tools covering almost all aspects of drug design: from biological target identification, receptor modeling, virtual screening, and QSAR analysis to absorption, metabolism, and toxicity prediction. It is difficult to assess how many of the currently registered drugs were developed using computational methods. Some of the most renown examples are dorzolamide,[1] zanamivir,[2] raltegravir,[3] and aliskiren.[4]

Ligand−protein docking programs are one of the most widely used tools in drug development. They can be used for identification of novel lead compounds[5−7] and preliminary lead optimization[8−10] and for investigating mechanisms of action of biologically active compounds.[11,12] The main application of docking tools, however, is predicting the correct geometry of a ligand−protein complex. Many reports of virtual screening of chemical databases,[6,13−15] comparisons of various docking methods and scoring functions,[16−21] or methods for improving screening hit rates[22−24] have been published. In many such studies, authors base their assessment mainly on the docked ligand geometry and its topology in the binding site (ligand pose). This approach has several limitations. First, ligand poses with the best geometries in terms of RMSD (with respect to a reference ligand) are not necessarily assigned with the best docking scores. Second, a docking protocol can find low RMSD poses for both 1 mM and 1 nM active compounds; therefore, virtual screening methods that are focused on finding valid binding geometry can return many poor binders along with highly active compounds. Third, computing RMSD to a reference ligand assumes that both the bound ligand and binding site residues lose every rotational and translational degree of freedom upon complex formation, and the only possible geometry is the one observed in the crystallographic structure. Finding a plausible ligand pose is only half of the success; all poses need to be subsequently scored and ranked by an appropriate function that takes into account various interaction terms between the ligand and the receptor, such as electrostatic and van der Waals forces, hydrogen bonding, water displacement, conformational entropy changes, etc.[25,26] Therefore, the docking accuracy should be considered in terms of both binding geometry and predicted binding energy. Because the interactions between a ligand and a protein contribute to the biological activity of the ligand, it becomes apparent that there must be some correlation between scoring function values and the actual biological activity, though in many cases this correlation can be masked by various factors. Numerous attempts were made in order to find such correlations with various success.[16−19,27,28] There are several computational techniques that can be used to more accurately assess interactions between a ligand and a protein, like the free energy perturbation method or potential of mean force, but they are much more computationally demanding than molecular docking. Therefore, the idea to quickly predict the biological activity of large groups of compounds with docking tools is very attractive.

Received: July 22, 2013
Published: November 22, 2013

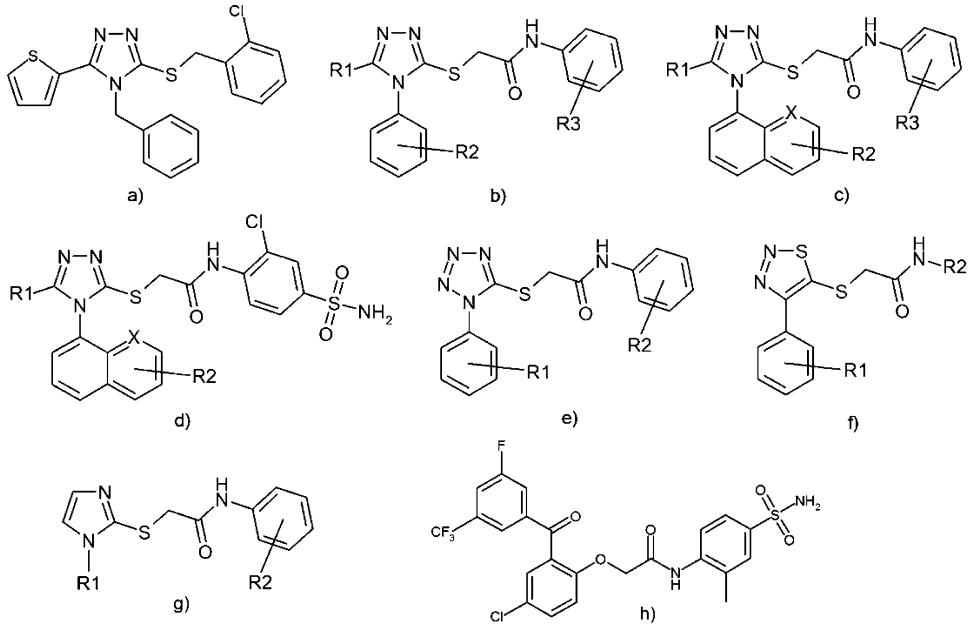


Figure 1. (a) Triazole NNRTI from 2RKI crystallographic structure. (b–g) General formulas of azole NNRTIs used in this study. (h) Benzophenone NNRTI from 3DLG crystallographic structure.

We have begun development of a novel computation-based workflow for rational drug design that features binding isotope effect studies. Our biological target is the human immunodeficiency virus reverse transcriptase (HIV-1 RT or RT). Reverse transcriptase is an enzyme crucial for HIV replication in the human host. There are several binding sites in the protein that can be used to inhibit its enzymatic activity. Our studies target the hydrophobic, allosteric, binding pocket called non-nucleoside inhibitors binding site (NNIBS), which is formed upon interaction of an inhibitor with the protein. In the first phase of our study, we identify new non-nucleoside inhibitors of reverse transcriptase (NNRTIs) with the 1,2,4-triazole core by the means of molecular docking. Bearing in mind that scoring functions of docking programs generally lack accuracy and their performance is highly dependent on the protein target,[19,29] this case study is aimed at assessing which docking algorithms and scoring functions perform best for the triazole NNRTIs-RT system.

Most of the docking tools return one or more values for ligand–receptor interactions in some arbitrary units. In order to predict biological activities using these values, appropriate correlation functions must be found. This in turn requires experimental activity values of good quality. There is no guarantee, however, that a simple correlation function (e.g., linear) can be obtained and that the function will be applicable to other ligand–receptor systems. There is a simple solution to circumvent this problem. Assuming that the correlation function exists and is monotonic, the relative activity of two compounds can be obtained simply by comparing their calculated values of interactions. Instead of predicting actual activity values, a simple classification of compounds as more or less active than a given reference is sufficient for most of the drug design process. The advantage of this approach is that by using rank correlation methods like Spearman's ρ coefficient, any linear or nonlinear correlation can be found if such exists. Moreover, rank correlation methods can perform reasonably well with low quality data. The nonlinearity and insensitivity to small errors in the input data makes the ρ coefficient more suitable for scoring functions assessment than the Pearson's linear coefficient, which is often used for this purpose.[16,17,19,25]

EVALUATION METHODOLOGY

Ligand Sets. To assess how accurately the scoring functions of the docking applications are for identifying the more active compound from two compounds, we prepared three test sets of 1,2,4-triazole-based NNRTIs. Choosing the experimental data is crucial for every study regarding quantitative structure–activity model building or validation/comparison of computational methods. The problem of obtaining a good quality set of consistent and accurate data is especially difficult in medicinal chemistry. Most of the available data on biological activity is published only for small sets of compounds. A multitude of methods for measuring biological activity makes it difficult to compare data from different sources. Moreover, there is no tendency to cross-validate experimental values from different laboratories; ideally, all measurements of biological activity should be performed in at least two independent laboratories. Given all this, it is very problematic to choose appropriate data for the analysis. When selecting ligand sets for this study, we focused mainly on obtaining large and consistent data on biological activity. The first set comprised 32 benzylsulfanyltriazoles described by Kirschberg et al. with pEC_{50} ranging from 4.29 to 6.88. All compounds in that set are closely related to 4-benzyl-3-[(2-chlorobenzyl)sulfanyl]-5-thien-2-yl-4H-1,2,4-triazole (Figure 1a), which is present in the crystallographic structure of HIV-RT published by the same authors (Protein Data Bank code 2RKI).[30]

Sets 2 and 3 consist of potent inhibitors (55 and 24 compounds, respectively) with a triazolylthioglycolanilide scaffold (Figure 1b–d).[31] All compounds from set 3 contain a sulfanilamide motif, which is very common in the second generation NNRTIs (Figure 1d).[32−36] Activities (pEC_{50}) for sets 2 and 3 range from 5.64 to 10.0. Because all 111 test compounds come from two publications only and the biological activity values for all of the compounds are reported as mean values of multiple experiments, our choice of test compounds

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

seemed to be consistent and of good quality. During our study, however, doubts arose about the accuracy of activity values for sets 2 and 3 (see Results and Discussion). For this reason, we needed to prepare an additional three sets of NNRTIs: 25 tetrazoles (set 4, Figure 1e),[32] 34 1,2,3-thiadiazoles (set 5, Figure 1f),[37,38] and 17 imidazoles (set 6, Figure 1g).[39] All the inhibitors in sets 2−6 contain the same aryl-azole-thioglycolamide core.

It should be noted that the EC_{50} value is affected by many factors other than ligand−receptor interactions; therefore, it cannot be treated as a direct binding affinity measure. However, assuming that all or most of the discussed compounds behave similarly in vivo (i.e., there are no drastic differences in interactions with other components of the biological system), the EC_{50} value can reflect the binding affinity. Because cell-based assays are one of the simplest biological systems, many factors that can interfere with ligand−RT interactions are eliminated. We have adopted this assumption because it would be much more difficult to compile a sufficient data set comprising only IC_{50} or K_i values.

Docking Programs. All 111 test compounds were docked using four docking programs: Glide,[40−42] Molegro Virtual Docker (MVD, formerly MolDock),[43] FlexX,[44] and AutoDock Vina.[45] Additionally, the Hyde scoring function[46] was used for the assessment of docked poses returned by Glide. For some of the programs, different settings and protocols were used in order to explore their performance in a triazole−RT system. The resulting poses were ranked according to the main scoring function values, their selected components, and alternative scoring functions (if present).

Receptors. Two crystallographic structures of RT were used for the preparation of receptors. The first protein with the Protein Data Bank (PDB) code 2RKI is the only triazole-containing NNRTI−RT complex deposited in PDB to date. The shape of its binding site is well suited for ligands from set 1, as mentioned before. Also, the resolution of 2.3 Å is acceptable. The second protein 3DLG (resolution 2.2 Å) contains a benzophenone inhibitor (GW69564, Figure 1h). The choice of this structure was dictated by several factors. The RMSD between active the site residues of 2RKI and 3DLG is 0.898, which is one of the lowest from all of the RT structures with resolution equal or less than 2.3 Å (Table S19, Supporting Information). The binding mode of the benzophenone ligand from 3DLG has the important feature of an aromatic substituent occupying the hydrophobic tunnel between Val106 and Pro236, which is also suspected for triazole NNRTIs (see below). Also, the benzophenone has the sulfanilamide substituent analogous to the sulfanilamide motif present in the ligands in set 3. Moreover, 3DLG was already used in several docking studies with aryl-azole-thioglycolanilide NNRTIs.[37,39,47]

Pose Assessment. For the optimal binding affinity prediction using scoring functions, it is crucial to choose the correct ligand pose. This is a difficult task because there are no crystallographic data for triazole NNRTIs, except for 2RKI. The supposed binding mode of azole NNRTIs is based mainly on molecular modeling. Several studies exist that discuss the binding of azole inhibitors to RT, and to the best of our knowledge, there is only one binding mode proposed for this group of compounds (Figure 2).[37,39,48−50] SAR studies also support this binding mode; the pocket formed by Glu138, Val179, and Ile180 is small; therefore, a larger substituent at the C5-position of the triazole ring leads to a decrease in activity.[30,50] Long and rigid chains can be added in the *para* positon of the anilide ring as it protrudes from the hydrophobic tunnel between Val106 and Pro236 that connects the binding site with the protein surface;[48] the *para* position of an aromatic substituent at N4 of the triazole ring is directed toward Trp229 (crucial for RT activity), adding a group that increases the hydrophobic contact with this amino acid enhances the inhibitory activity.[31,37,38] The core structure of compounds from sets 2−6 is basically the same. For set 1 there are some differences, but the assumed binding mode is very similar.

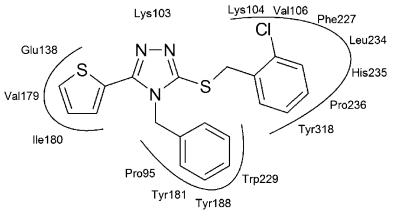


Figure 2. Schematic representation of triazole NNRTI orientation in the NNIBS. Three hydrophobic pockets are occupied by substituents of triazole core.

All ligands poses resulting from docking were visually inspected. The poses that matched the assumed binding mode (Figure 2) were considered valid and put to the separate set (Valid Poses). Because it cannot be excluded that the different binding modes found during docking were in fact correct for some of the inhibitors, the best poses returned for each docking protocol were also used for activity prediction without the discrimination of binding geometries (All Poses). Therefore the term "valid pose" is used throughout the paper to refer to a pose that is in agreement with previously reported binding mode and does not mean that the pose is necessarily correct.

Receptor Flexibility. The HIV-RT is a flexible protein, and different chemical groups of NNRTIs are known to bind with the enzyme in several distinct conformations. For this reason, the binding site flexibility should be taken into account in order to correctly dock the ligands. There are several approaches that address this problem: ensemble docking,[51,52] side chain rotamers,[53,54] Monte Carlo minimization,[55] molecular dynamics,[56,57] normal modes analysis,[53,58] and others.[59,60] While most of these methods achieved some success, none of them is considered as the universal solution to the receptor flexibility problem. To deal with the RT flexibility in our study, the following steps were undertaken. First, as described above, we assumed that there is only one valid binding mode for all of the used ligands; therefore, the conformational changes of the receptor are expected to be limited. Second, two different receptor conformations were used (the idea of ensemble docking) that we considered to be the most similar to the actual binding mode of azole NNRTIs based on our experience and literature data. Third, we used softened van der Waals potentials for ligand and/or receptor atoms (FlexX, MVD, Glide) to simulate a small scale relaxation of the ligand−receptor complex. Finally, we employed two docking protocols that include the receptor changes: flexible residues in MVD (failed) and Induced Fit Docking[61,62] (successful).

Statistical analysis. For the statistical analysis Spearman's ρ coefficient (r_s) was calculated according to the formula

$$r_s = \frac{\frac{1}{6}(n^3 - n) - (\sum_{i=1}^{n} d_i^2) - T_X - T_Y}{\sqrt{\left(\frac{1}{6}(n^3 - n) - 2T_X\right)\left(\frac{1}{6}(n^3 - n) - 2T_Y\right)}} \quad (1)$$

where

$$d_i = Rx_i - Ry_i \quad (2)$$

$$T_X = \frac{1}{12}\sum_j (t_j^3 - t_j) \quad (3)$$

$$T_Y = \frac{1}{12}\sum_k (u_k^3 - u_k) \quad (4)$$

Rx_i and Ry_i are ranks for literature data and docking score classifications respectively, t_j is number of samples with j-th rank value for literature data, u_k is number of samples with k-th rank value for docking score, n is number of samples.[63]

In order to check if a blocking of the biological activity values can increase the correlation for set 2 and 3 (see Results and Discussion), we used a modified formula for the ρ coefficient

$$r_s = 1 - \frac{6(\sum_{i=1}^{n} d_i^2 + T_X + T_Y)}{n^3 - n} \quad (5)$$

Equation 5 is appropriate when tied ranks that result from data aggregation (blocking) are present. Test of statistical significance was carried out using the following statistic:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \quad (6)$$

which distribution is approximated by Student's t-distribution with n-2 degrees of freedom.

RESULTS AND DISCUSSION

Null Model. For the comparison with the results obtained from docking, we created a null model for each data set. The Spearman's ρ coefficients were calculated for molar masses and pEC_{50} values of each ligand set. For sets 1−3, the correlations coefficients are, respectively, −0.582 (p-value: 4.75e-04), −0.211 (p-value: 1.22e-01), and 0.352 (p-value: 9.14e-02). Significant correlation can be observed only for set 1. This result can be explained by the fact that for 32 compounds in set 1 there are only 12 different molecular masses. As shown in Figure 3, many compounds with significantly different biological activities have the same mass; for example, there are 9 compounds with mass 432.39 with pEC50 ranging from 5.05 to 6.14, and 6 compounds with mass 397.94, including the most and the least active compound in the set. Because many extreme data points were assigned with the same rank, the correlation coefficient was overestimated and should not be considered. For the three additional azole sets, the correlation values are −0.172 (p-value: 4.12e-01) for set 4, −0.600 (p-value: 1.75e-04) for set 5, and −0.654 (p-value: 4.44e-03) for set 6. Only for the 1,2,3-thiadiazoles (set 5) and imidazoles (set 6) is significant correlation observed. The correlation in set 5 is a result of increasing activity of congeneric subseries of compounds where the R1 substituents are different halogens (Table S5, Supporting Information). With the increasing mass and volume of the halogen atoms, the hydrophobic contact with the binding pocket formed by Tyr188, Phe227, and Trp229 also increases, which correlates with the biological activity.[37,38]

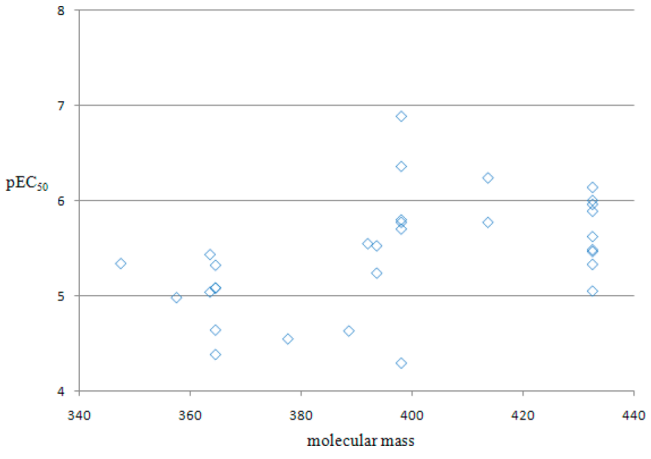


Figure 3. Null model for set 1.

Glide SP. Selected Spearman's ρ coefficients (r_s) for ligands ranked on the basis of Glide SP (Standard Precision) docking results are presented in Table 1. The complete data is available in the Supporting Information. The p-values are given in parentheses. The valid poses [%] value represents the ratio of ligands with at least one valid pose (among the three top scoring poses) to all of the ligands in a given set.

For 2RKI in combined sets 1, 2, and 3, the r_s values are not significantly different from 0 (random classification) except for *ecoul*, which has r_s = 0.576 for valid geometry, and *emodel* with negative r_s. On the contrary, for 3DLG, almost all of the classifications show some correlation—from 0.267 for *evdw* to 0.592 for *ecoul*. If we look at the results for each set separately, we see that for both 2RKI and 3DLG classifications are random for all of the values in sets 2 and 3. Good results in set 1 for 2RKI can be easily explained because compounds in set 1 have similar structure to the reference ligand, thus they can adapt to the geometry of the binding site of 2RKI. Interestingly, also for 3DLG, the best results are obtained for set 1 even though there is almost no similarity in the chemical structure between compounds in set 1 and the 3DLG reference ligand. The results agree with the assumption that both triazoles and benzophenones adapt similar geometry in the NNIBS, most likely because the nonspecific hydrophobic interactions play a key role in the binding of NNRTIs.

For 2RKI, Glide SP gives a good prediction of ligand orientation in the binding site, reaching 100% correctness for sets 1 and 2. For 3DLG, the prediction is worse but still among the best of all docking protocols used in this study, with the highest value for set 2 (92.7%). The rise of r_s for ligands with plausible binding geometry is apparent for set 1 (3DLG).

Glide XP. Although Glide XP (Extra Precision) mode is only a different protocol of Glide program, it is essentially a new docking method, which relies not only on a conformational search but attempts to regrow and optimize the ligand in the binding site. To assess the obtained poses, a special scoring function is applied, which includes several additional interaction terms, such as desolvatation effects, more accurate treatment of hydrophobic interactions, and various hydrogen bonds categories.[64] Glide XP performs more accurate geometrical sampling of the ligand; therefore, it needs significantly more time than the SP mode (519.7 s for 1 ligand per 1 CPU core on the average, as compared to Glide SP that needed 14.5 s for 1 ligand per core of Xeon X5460 processor). Table 2 contains results for the Glide XP mode. Notably, there is a significant

Table 1. Glide SP Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			97.3	**all sets**			82.9
gscore[a]	−0.054 (5.75e-01)	−0.019 (8.48e-01)		*gscore*	0.278 (3.09e-03)	0.421 (2.91e-05)	
emodel[b]	−0.205 (3.19e-02)	−0.203 (3.59e-02)		*emodel*	0.360 (1.04e-04)	0.423 (2.69e-05)	
ecoul[c]	0.548 (6.01e-10)	0.576 (8.43e-11)		*evdw*[d]	0.267 (4.61e-03)	0.317 (2.08e-03)	
				ecoul	0.592 (8.03e-12)	0.505 (2.79e-07)	
				energy[e]	0.385 (2.95e-05)	0.411 (4.80e-05)	
set 1			100	**set 1**			75.0
gscore	0.343 (5.49e-02)	0.347 (5.20e-02)		*gscore*	−0.088 (3.59e-01)	−0.196 (6.18e-02)	
emodel	0.559 (8.87e-04)	0.559 (8.87e-04)		*emodel*	0.395 (2.51e-02)	0.539 (6.58e-03)	
evdw	0.553 (1.02e-03)	0.553 (1.02e-03)		*evdw*	0.293 (1.04e-01)	0.408 (4.78e-02)	
energy	0.542 (1.35e-03)	0.542 (1.35e-03)		*energy*	0.457 (8.51e-03)	0.506 (1.16e-02)	
				eint[f]	−0.294 (1.03e-01)	−0.536 (6.90e-03)	
set 2			100	**set 2**			92.7
gscore	0.106 (4.45e-01)	0.124 (3.73e-01)		*gscore*	0.115 (4.05e-01)	0.134 (3.50e-01)	
set 3			87.5	**set 3**			70.8
gscore	−0.193 (3.65e-01)	−0.194 (4.01e-01)		*gscore*	0.119 (5.79e-01)	0.102 (6.96e-01)	

[a]*gscore*: GlideScore. [b]*emodel*: model energy. [c]*ecoul*: Coulomb energy. [d]*evdw*: van der Waals energy. [e]*energy*: modified Coulomb−van der Waals interaction energy. [f]*eint*: internal torsional energy.

Table 2. Glide XP Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			59.5	**all sets**			67.6
gscore[a]	0.042 (6.65e-01)	0.240 (5.21e-02)		*gscore*	0.293 (1.82e-03)	0.379 (8.01e-04)	
emodel[b]	−0.060 (5.32e-01)	−0.123 (3.23e-01)		*emodel*	0.583 (1.95e-11)	0.192 (9.90e-02)	
evdw[c]	0.232 (1.42e-02)	0.024 (8.46e-01)		*evdw*	0.550 (4.16e-10)	0.258 (2.56e-02)	
ecoul[d]	0.286 (2.30e-03)	0.272 (2.72e-02)		*ecoul*	0.510 (1.11e-08)	0.091 (4.37e-01)	
energy[e]	0.282 (2.70e-03)	0.104 (4.04e-01)		*energy*	0.622 (3.22e-13)	0.232 (4.51e-02)	
set 1			53.1	**set 1**			3.1
gscore	0.321 (7.30e-02)	0.604 (1.03e-02)		*gscore*	−0.018 (9.21e-01)	–	
emodel	0.356 (4.57e-02)	0.600 (1.09e-02)		*eint*[f]	−0.497 (3.79e-03)	–	
evdw	0.200 (2.73e-01)	0.730 (8.77e-04)					
energy	0.138 (4.52e-01)	0.704 (1.60e-03)					
eint	−0.389 (2.78e-02)	−0.190 (4.65e-01)					
set 2			67.3	**set 2**			94.5
gscore	0.286 (3.42e-02)	0.301 (7.08e-02)		*gscore*	0.275 (4.18e-02)	0.299 (3.12e-02)	
				evdw	0.198 (1.47e-01)	0.220 (1.17e-01)	
				energy	0.135 (3.27e-01)	0.140 (3.22e-01)	
set 3			50.0	**set 3**			91.7
gscore	0.258 (2.24e-01)	0.197 (5.40e-01)		*gscore*	0.456 (2.50e-02)	0.517 (1.36e-02)	
				evdw	0.106 (6.23e-01)	0.090 (6.90e-01)	
				energy	0.110 (6.10e-01)	0.066 (7.71e-01)	

[a]*gscore*: GlideScore. [b]*emodel*: model energy. [c]*ecoul*: Coulomb energy. [d]*evdw*: van der Waals energy. [e]*energy*: modified Coulomb−van der Waals interaction energy. [f]*eint*: internal torsional energy.

drop in the valid poses prediction rate as compared to the SP mode. This is mainly because Glide XP returned only one pose for almost all of the docked compounds, which means that the differences between the best poses were below a set threshold. The r_s values for 2RKI in combined sets and in sets 2 and 3 are low and not statistically significant. Similarly to the SP mode, the best correlation is obtained for set 1: from 0.600 for *emodel* to 0.730 for *evdw*. Four of six values returned by Glide XP for set 1 have r_s = 0.600 or more, which is an apparent improvement as compared to Glide SP. If no discrimination between valid and invalid poses of the ligand is made, most of the r_s values for set 1 are significantly lower. Therefore, finding a plausible ligand pose is crucial for scoring using Glide XP. Contradictory conclusions can be reached at first when examining the results for the combined sets for 3DLG: four values have r_s greater than 0.500 for all poses, which drops significantly for valid poses. Inspection of the results for sets 1 through 3 separately indicates that the correlations for all poses and for valid poses are similar. The reason for the drop of correlation for valid poses in the combined sets is the failure of the docking algorithm to find correct poses for all but one of the ligands in set 1. Another result that may appear strange at first glance is that the *emodel*, *evdw*, *ecoul*, and *energy* values have significant r_s values in combined sets (all poses, 3DLG), but no significant correlation is observed for separate sets. This results from scoring most of the compounds from sets 2 and 3 higher than set 1, comprising less active inhibitors (Figure S5, Supporting Information). Possibly, the scoring functions

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

Table 3. Induced Fit Docking Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			97.3	**all sets**			88.3
gscore[a]	0.383 (3.32e-05)	0.308 (1.18e-03)		*gscore*	0.366 (7.84e-05)	0.453 (2.90e-06)	
emodel[b]	0.337 (2.94e-04)	0.224 (1.96e-02)		*emodel*	0.592 (7.98e-12)	0.426 (1.25e-05)	
evdw[c]	0.377 (4.59e-05)	0.279 (3.52e-03)		*evdw*	0.487 (6.13e-08)	0.313 (1.70e-03)	
ecoul[d]	0.476 (1.28e-07)	0.447 (1.20e-06)		*ecoul*	0.622 (3.25e-13)	0.574 (6.46e-10)	
energy[e]	0.429 (2.62e-06)	0.354 (1.70e-04)		*energy*	0.588 (1.15e-11)	0.435 (7.70e-06)	
eint[f]	0.201 (3.40e-02)	0.077 (4.25e-01)		*eint*	0.153 (1.08e-01)	0.354 (3.45e-04)	
ifdscore[g]	0.343 (2.25e-04)	0.261 (6.27e-03)		*ifdscore*	0.355 (1.33e-04)	0.404 (3.61e-05)	
set 1			100	**set 1**			62.5
gscore	0.125 (4.95e-01)	0.275 (1.28e-01)		*gscore*	0.284 (1.15e-01)	0.432 (5.68e-02)	
emodel	0.694 (1.04e-05)	0.700 (8.36e-06)		*emodel*	0.495 (4.00e-03)	0.474 (3.48e-02)	
evdw	0.637 (8.77e-05)	0.637 (8.77e-05)		*evdw*	0.512 (2.71e-03)	0.366 (1.13e-01)	
energy	0.611 (2.03e-04)	0.615 (1.78e-04)		*energy*	0.498 (3.70e-03)	0.322 (1.66e-01)	
				ifdscore	0.197 (2.81e-01)	0.46 (4.15e-02)	
set 2			94.5	**set 2**			100
gscore	0.294 (2.92e-02)	0.235 (9.33e-02)		*gscore*	0.230 (9.06e-02)	0.190 (1.64e-01)	
evdw	0.354 (8.03e-03)	0.221 (1.15e-01)					
energy	0.285 (3.49e-02)	0.198 (1.60e-01)					
ifdscore	0.269 (4.73e-02)	0.117 (4.08e-01)					
set 3			100	**set 3**			95.8
gscore	−0.051 (8.13e-01)	0.006 (9.79e-01)		*gscore*	0.008 (9.71e-01)	−0.077 (7.27e-01)	
				eint	0.458 (2.44e-02)	0.368 (8.42e-02)	

[a]*gscore*: GlideScore. [b]*emodel*: model energy. [c]*ecoul*: Coulomb energy. [d]*evdw:* van der Waals energy. [e]*energy*: modified Coulomb−van der Waals interaction energy. [f]*eint*: internal torsional energy. [g]*ifdscore*: GlideScore + 0.05 × Prime energy.

Table 4. AutoDock Vina Results

	2RKI				3DLG		
	r_s all poses	r_s valid poses	valid poses [%]		r_s all poses	r_s valid poses	valid poses [%]
all sets			94.6	**all sets**			75.7
score	−0.628 (1.78e-13)	−0.652 (5.18e-14)		*score*	0.181 (5.68e-02)	0.259 (1.74e-02)	
set 1			100	**set 1**			28.1
score	0.522 (2.17e-03)	0.522 (2.17e-03)		*score*	0.425 (1.54e-02)	−0.508 (1.62e-01)	
set 2			90.9	**set 2**			92.7
score	−0.174 (2.03e-01)	−0.217 (1.30e-01)		*score*	−0.104 (4.51e-01)	−0.063 (6.62e-01)	
set 3			95.8	**set 3**			100
score	0.295 (1.62e-01)	0.220 (3.13e-01)		*score*	0.118 (5.82e-01)	0.114 (5.95e-01)	

could not correctly predict the effect of small chemical modifications of the core structure on biological activity. They could, however, identify the more active compound from two ligands with different chemical scaffolds or with significant differences in activity. Similar behavior was observed for some of the other docking methods used. Notably, *gscore* for set 3 in 3DLG attained r_s = 0.517 for valid poses, which is one of the highest values for sets 2 and 3 for any scoring method used in this study. This in turn suggests that activity of compounds from set 3 might be determined by some specific interactions included in XP mode *gscore* that are not recognized by other scoring functions.

Induced Fit Docking (IFD). Maestro suite offers a protocol for modeling ligand-induced binding site alterations by combining Glide and Prime (a program for predicting protein structures). The results obtained with this protocol are in many ways analogous to other Glide protocols (Table 3). There is, however, a substantial improvement in valid pose prediction. It should be noted that IFD returns several receptor conformations, and selecting the best solutions is not trivial because not only the docking score for ligand should be taken into account but also the energy and quality of the predicted receptor conformation. Therefore, valid pose percentages might be overestimated because the ligand orientations were selected not from 3 but from around 10 returned poses.

The absolute r_s values for most of the classifications decrease for poses that were considered plausible. This suggest that for IFD simple classification of poses as valid or invalid according to rough visual orientation assessment (Figure 2) is not sufficient. Therefore a more sophisticated discrimination method should be used that also includes receptor conformation.

For an ideal docking method that can correctly predict conformational changes of a receptor, results for a given ligand should be identical regardless of the starting receptor conformation. IFD, however, is dependent on receptor input structures because the first step of this protocol is docking to a rigid receptor. Therefore results for 2RKI and 3DLG are not identical but similar to corresponding rigid protein docking (Glide XP). Analogous to other Glide based docking, r_s values for combined ligand sets are higher for 3DLG than for 2RKI. Also, the best correlations are achieved in set 1, while in sets 2 and 3 almost no statistically significant correlation is observed. Because no significant improvement of classification rates was

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

Table 5. FlexX Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			65.8	**all sets**			36.0
score[a]	−0.045 (6.41e-01)	0.006 (9.59e-01)		*score*	0.635 (7.59e-14)	0.656 (4.37e-06)	
match[b]	0.262 (5.50e-03)	0.344 (2.91e-03)		*match*	0.582 (2.17e-11)	0.655 (4.50e-06)	
				lipo[c]	0.031 (7.50e-01)	0.504 (9.15e-04)	
set 1			100	**set 1**			34.4
score	0.414 (1.86e-02)	0.428 (1.46e-02)		*score*	−0.175 (3.39e-01)	0.491 (1.25e-01)	
lipo	0.528 (1.90e-03)	0.538 (1.51e-03)					
set 2			34.5	**set 2**			10.9
score	0.094 (4.94e-01)	0.012 (9.60e-01)		*score*	0.047 (7.31e-01)	−0.086 (8.72e-01)	
set 3			91.7	**set 3**			95.8
score	−0.152 (4.78e-01)	−0.229 (3.05e-01)		*score*	0.184 (3.90e-01)	0.152 (4.89e-01)	

[a]*score*: FlexX docking score. [b]*match*: matched interacting groups. [c]*lipo*: lipophilic contact area.

Table 6. Molegro Virtual Docker Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			61.3	**all sets**			71.2
score[a]	−0.037 (7.00e-01)	−0.056 (6.50e-01)		*score*	0.013 (8.96e-01)	0.031 (7.85e-01)	
rerank[b]	−0.416 (5.70e-06)	−0.663 (7.57e-10)		*rerank*	0.002 (9.82e-01)	0.096 (3.99e-01)	
set 1			96.9	**set 1**			59.4
score	0.426 (1.51e-02)	0.358 (4.80e-02)		*score*	0.410 (1.99e-02)	0.467 (4.40e-02)	
rerank	0.336 (6.02e-02)	0.289 (1.15e-01)		*rerank*	0.016 (9.31e-01)	0.074 (7.64e-01)	
set 2			36.4	**set 2**			76.4
score	0.092 (5.05e-01)	−0.008 (9.72e-01)		*score*	−0.013 (9.27e-01)	−0.034 (8.32e-01)	
rerank	−0.165 (2.29e-01)	−0.165 (4.88e-01)		*rerank*	−0.200 (1.43e-01)	−0.131 (4.07e-01)	
set 3			70.8	**set 3**			75.0
score	0.007 (9.74e-01)	−0.079 (7.63e-01)		*score*	−0.199 (3.51e-01)	−0.322 (1.93e-01)	
rerank	0.420 (4.12e-02)	0.073 (7.81e-01)		*rerank*	−0.089 (6.79e-01)	−0.430 (7.51e-02)	

[a]*score*: MolDock Score [Grid]. [b]*rerank*: default rerank scoring function.

obtained for IFD, it supports the assumption that there are no large changes in the NNIBS conformation upon binding of different triazoles.

The IFD protocol combines two Glide docking calculations and one Prime refinement; therefore, it needs substantially more time to complete (18216.6 s on average for one ligand per one core of Xeon X5460 processor).

AutoDock Vina. Spearman's ρ coefficients for Vina scores are presented in Table 4. High negative correlation for combined sets docked to 2RKI can be observed. Analogous to other docking protocols in this study, a significant correlation was obtained only for set 1. All ligand poses from set 1 were very similar to the pose of the reference ligand from the crystal structure. Vina scoring function apparently favored a good fit of ligands in set 1 and assigned them with the higher scores. Therefore, set 1 was positioned above sets 2 and 3, which resulted in a negative correlation for the combined sets (Figure S6, Supporting Information). For 3DLG, the program performance was worse. There is a strange correlation shift from positive to negative in set 1 when invalid poses are removed. It is most likely a result of insufficient accurate discrimination between valid and invalid ligand poses or the scoring function's fault. Like other docking programs, Vina was unable to correctly rank inhibitors from sets 2 and 3. It is worth noting that the program returned consistent poses with plausible orientation in the binding site for most of the highest ranked compounds. Vina's scoring function high dependency on used receptor conformation limits this program's use for biological activity prediction.

FlexX. The results for FlexX are not very different from those obtained for the rest of docking protocols (Table 5). For 2RKI, the best r_s values are achieved in set 1 by main score and its lipophilic interactions component. The outcome for 3DLG is interesting. There is a good correlation for all scores of valid poses in combined sets, while no statistically significant correlation in separate sets can be seen (although the r_s values are higher for set 1 than for sets 2 and 3). This is probably because even though there is a low correlation in each set, the relative positioning of compounds from different sets result in a higher correlation for combined sets. This explanation, however, is credible only for valid poses. It is rather surprising that *FlexX score* and *match* have high r_s values for unfiltered results given the low valid pose prediction rate (36.0%). It is possible that despite improper orientation of some ligands, their binding mode was correlated with the biological activity. FlexX achieved the worst valid pose prediction rate for 3DLG from all programs. A good rate was obtained only for set 3 probably due to formation of hydrogen bonds by the sulfonamide substituents, which facilitated correct orientation of the ligands.

Molegro Virtual Docker. The docking with MVD was performed using two different partial charge schemes for both proteins and ligands(Table 6). First were the OPLS-2005[65] charges assigned by LigPrep utility[66] (for the ligands) and Protein Preparation Wizard (for the proteins).[67,68] The second

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

Table 7. Molegro Virtual Docker with "Flexible Residues" Results

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
all sets			73.0	**all sets**			64.0
score[a]	0.049 (6.11e-01)	0.024 (8.31e-01)		*score*	0.088 (3.57e-01)	0.097 (4.23e-01)	
rerank[b]	−0.418 (4.91e-06)	−0.663 (1.57e-11)		*rerank*	0.010 (9.15e-01)	0.000 (9.99e-01)	
set 1			93.8	**set 1**			31.3
score	0.638 (8.55e-05)	0.607 (3.78e-04)		*score*	0.368 (3.82e-02)	−0.176 (6.27e-01)	
rerank	0.548 (1.18e-03)	0.506 (4.29e-03)		*rerank*	0.028 (8.78e-01)	−0.091 (8.03e-01)	
set 2			54.5	**set 2**			72.7
score	0.089 (5.17e-01)	−0.042 (8.26e-01)		*score*	−0.037 (7.87e-01)	−0.010 (9.51e-01)	
rerank	−0.200 (1.43e-01)	−0.434 (1.67e-02)		*rerank*	−0.213 (1.18e-01)	−0.311 (5.05e-02)	
set 3			87.5	**set 3**			87.5
score	0.013 (9.52e-01)	−0.067 (7.75e-01)		*score*	0.150 (4.85e-01)	0.151 (5.12e-01)	
rerank	0.144 (5.03e-01)	−0.082 (7.25e-01)		*rerank*	0.241 (2.57e-01)	0.142 (5.40e-01)	

[a]*score*: MolDock Score [Grid]. [b]*rerank*: default rerank scoring function.

Table 8. Hyde Results

	2RKI			3DLG	
scoring method	r_s	r_s (no Hyde optimization)	scoring method	r_s	r_s (no Hyde optimization)
all sets			**all sets**		
Hyde score	−0.522 (4.92e-09)	−0.512 (1.05e-08)	*Hyde score*	0.231 (1.46e-02)	0.164 (8.64e-02)
ΔG desolv[a]	−0.483 (9.11e-08)	−0.512 (1.08e-08)	*ΔG desolv*	−0.468 (2.22e-07)	−0.359 (1.11e-04)
ligand E desolv[b]	−0.574 (5.43e-11)	−0.588 (1.49e-11)	*ligand E desolv*	−0.286 (2.34e-03)	−0.130 (1.74e-01)
ligand lipo[c]	0.178 (6.22e-02)	0.23 (1.58e-02)	*ligand lipo*	0.262 (5.52e-03)	0.228 (1.63e-02)
set 1			**set 1**		
Hyde score	0.506 (3.12e-03)	0.524 (2.08e-03)	*Hyde score*	0.219 (2.29e-01)	−0.002 (9.91e-01)
ΔG desolv	0.495 (3.97e-03)	0.524 (2.08e-03)	*ligand E desolv*	0.571 (6.41e-04)	0.435 (1.29e-02)
ligand E desolv	0.548 (1.16e-03)	0.546 (1.22e-03)	*ligand lipo*	0.589 (3.91e-04)	0.485 (4.95e-03)
ligand lipo	0.523 (2.14e-03)	0.558 (9.12e-04)			
set 2			**set 2**		
Hyde score	0.179 (1.96e-01)	0.287 (3.52e-02)	*Hyde score*	0.110 (4.25e-01)	0.194 (1.55e-01)
ΔG desolv	0.261 (5.62e-02)	0.279 (4.10e-02)	*ΔG desolv*	0.181 (1.87e-01)	0.419 (1.43e-03)
ligand lipo	0.242 (7.73e-02)	0.287 (3.54e-02)	*ligand lipo*	0.324 (1.59e-02)	0.223 (1.01e-01)
set 3			**set 3**		
Hyde score	0.065 (7.62e-01)	−0.031 (8.85e-01)	*Hyde score*	0.585 (2.67e-03)	0.288 (1.73e-01)

[a]*ΔG desolv*: desolvatation energy. [b]*ligand E desolv*: ligand energy of desolvatation. [c]*ligand lipo*: ligand lipophilic contact.

were partial charges assigned by MVD, which are equal to zero for all neutral atoms. Because the used ligands are not ionized in pH close to neutral, using the MVD charges reduced the ligand−protein electrostatic interactions to zero, but docking results were slightly better than for the OPLS-2005 charges (Supporting Information). MVD is one of the docking programs that by default use a simplified charges scheme. The scoring functions of docking programs are parametrized on the basis of many ligand−protein complexes and perform optimally with original settings. Thus, it is not surprising that using other charges (e.g., partial atomic charges) often leads to a lowered docking accuracy. This also suggests that the steric and hydrophobic interactions are the most important for binding with NNIBS of RT.

MVD offers the possibility to model the conformational changes of a receptor upon ligand binding by optimizing the position of selected amino acids residues. This feature, however, is not as powerful of a tool as IFD, and by defining as little as five amino acids closest to the reference ligand, no results were obtained. Therefore to account for limited flexibility of the binding site, we reduced the steric interactions for selected amino acid residues (treated as "flexible") as described in the Methods section. For comparison we also performed docking to the rigid receptor (Table 7).

The *rerank* scores in the combined sets for 2RKI have significant negative r_s values for both rigid and "flexible" docking. On the other hand, both *rerank* and *MolDock* scores have fairly high positive r_s values in set 1. In sets 2 and 3, no statistically significant correlation is observed with two exceptions: *rerank* in set 2 for "flexible" docking with r_s = −0.434, and *rerank* for set 3 (rigid docking, all poses) with r_s = 0.420, which are among the highest correlations for sets 2 and 3 obtained in this work. Similarly to AutoDock Vina, the scoring function of MVD favors the high shape complementarity of compounds from set 1 and gives them better scores than for sets 2 and 3, hence, the negative r_s value for combined sets. For 3DLG, only rigid docking in set 1 gave some positive correlation (*score*). Also *rerank* in set 3 gave a relatively high negative correlation (p-value = 0.075). As expected, scaling down the interactions with the nearest amino acid residues increased the valid pose prediction rate for 2RKI and also slightly improved the ligand ranking success. The opposite effect was observed for 3DLG. Because this receptor has a larger volume of the binding site, reducing amino acids potentials lowered the docking accuracy.

Hyde. Because Hyde is a scoring function, it requires the ligands docked to the receptor as an input. Only the valid poses from Glide SP were used because this docking protocol had the

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

Table 9. Glide SP Results for Sets 4−6

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
set 4			76.0	**set 4**			76.0
gscore[a]	−0.027 (8.98e-01)	−0.036 (8.83e-01)		*gscore*	0.026 (9.04e-01)	0.278 (2.49e-01)	
emodel[b]	0.236 (2.56e-01)	0.314 (1.90e-01)					
evdw[c]	0.381 (6.05e-02)	0.688 (1.14e-03)					
energy[d]	0.245 (2.38e-01)	0.509 (2.59e-02)					
set 5			58.8	**set 5**			97.1
gscore	−0.054 (7.64e-01)	−0.550 (1.20e-02)		*gscore*	0.166 (3.48e-01)	0.276 (1.20e-01)	
evdw	0.264 (1.31e-01)	−0.001 (9.97e-01)		*emodel*	0.349 (4.33e-02)	0.397 (2.23e-02)	
ecoul[e]	−0.228 (1.94e-01)	−0.213 (3.66e-01)		*evdw*	0.464 (5.66e-03)	0.399 (2.13e-02)	
				energy	0.394 (2.11e-02)	0.397 (2.23e-02)	
set 6			88.2	**set 6**			100
gscore	0.701 (1.71e-03)	0.657 (7.78e-03)		*gscore*	0.825 (4.64e-05)	0.825 (4.64e-05)	
emodel	0.714 (1.30e-03)	0.868 (2.77e-05)		*emodel*	0.874 (4.50e-06)	0.874 (4.50e-06)	
evdw	0.448 (7.12e-02)	0.958 (2.08e-08)		*evdw*	0.763 (3.67e-04)	0.763 (3.67e-04)	
energy	0.636 (6.08e-03)	0.940 (2.00e-07)		*energy*	0.699 (1.80e-03)	0.699 (1.80e-03)	

[a]*gscore*: GlideScore. [b]*emodel*: model energy. [c]*evdw*: van der Waals energy. [d]*energy*: modified Coulomb−van der Waals interaction energy. [e]*ecoul*: Coulomb energy.

Table 10. Glide XP Results for Sets 4−6

	2RKI				3DLG		
scoring method	r_s all poses	r_s valid poses	valid poses [%]	scoring method	r_s all poses	r_s valid poses	valid poses [%]
set 4			92.0	**set 4**			84.0
gscore[a]	−0.164 (4.32e-01)	−0.100 (6.49e-01)		*gscore*	−0.124 (5.54e-01)	−0.029 (8.99e-01)	
evdw[b]	0.193 (3.55e-01)	0.374 (7.88e-02)					
set 5			23.5	**set 5**			97.1
gscore	0.054 (7.60e-01)	0.488 (2.20e-01)		*gscore*	0.335 (5.28e-02)	0.324 (6.57e-02)	
ecoul[c]	−0.410 (1.60e-02)	−0.390 (3.39e-01)		*emodel*	0.515 (1.81e-03)	0.509 (2.49e-03)	
				evdw	0.574 (3.89e-04)	0.570 (5.29e-04)	
				ecoul	0.406 (1.72e-02)	0.404 (1.97e-02)	
				energy	0.573 (3.94e-04)	0.577 (4.40e-04)	
set 6			64.7	**set 6**			100
gscore	0.559 (1.96e-02)	0.545 (8.72e-02)		*gscore*	0.661 (3.90e-03)	0.661 (3.90e-03)	
emodel[d]	0.551 (2.20e-02)	0.449 (1.71e-01)		*emodel*	0.709 (1.45e-03)	0.709 (1.45e-03)	
evdw	0.774 (2.65e-04)	0.943 (2.85e-05)		*evdw*	0.751 (5.17e-04)	0.751 (5.17e-04)	
energy[e]	0.849 (1.60e-05)	0.934 (5.14e-05)		*energy*	0.690 (2.17e-03)	0.690 (2.17e-03)	
eint[f]	−0.256 (3.22e-01)	−0.581 (6.45e-02)					

[a]*gscore*: GlideScore. [b]*evdw*: van der Waals energy. [c]*ecoul*: Coulomb energy. [d]*emodel*: model energy. [e]*energy*: modified Coulomb−van der Waals interaction energy. [f]*eint*: internal torsional energy.

highest valid pose prediction rate for both receptors (except for the IFD, but as mentioned before, it is difficult to decide which poses are valid for this protocol). By default, Hyde performs a pose energy minimization before scoring. Because the minimization step was already carried out by Glide, for the comparison, scoring was performed with and without the optimization (Table 8). Disabling the optimization slightly improved the results for sets 1, 2, and combined sets 1−3 for 2RKI, while for 3DLG the opposite was observed. Most of the scores for 2RKI in combined sets are negatively correlated with the relative activity, which results from higher scores assigned to ligands from set 1. Consistent with other scoring functions, the highest correlations are obtained for set 1. Interestingly, in set 2, low but statistically significant correlations are observed for three out of four scores without optimization. For 3DLG in combined sets, the results appear more random with some of the scores having tendency for negative correlation and others for positive. Notably, one or two scores in each set have r_s values over 0.400. In set 3, *Hyde score* for optimized poses obtained r_s = 0.585, which is the highest result for set 3 of all used scoring functions. Unfortunately, Hyde performance is not consistent; the obtained correlations vary depending on the used receptor, optimization method, ligand set, and scoring term. Therefore, it is difficult to point to which approach would give the most reliable biological activity predictions of triazole NNRTIs, which limits Hyde's use for this purpose.

Explaining the Poor Correlation in Sets 2 and 3. Several docking protocols were used in this study, each of them using a different search algorithm and scoring function. Even though those docking methods use various computational approaches, qualitative results obtained in this study are very consistent. However, the fact that none of the used docking tools were able to make accurate and confident relative activity predictions in two of three test sets questions the usefulness of the docking for that purpose. Therefore, it is necessary to explain the reason behind the poor performance of scoring functions in sets 2 and 3. There are several possible reasons for this. The simplest explanation is that the biological activity

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

values for those two sets of compounds are not correlated with the binding affinity. There are, however, some cases of statistically significant correlation for those sets in our study. Also, the correlation was found for some of the closely related group of compounds (see below). This, and some other indications described below suggest that the correlation may exist for sets 2 and 3, but it can be more difficult to find. Another reason might be the conformational flexibility of the reverse transcriptase, which should be taken into account when looking for the relation between binding affinity and biological activity. Because NNRTIs bind to the allosteric site of RT, their activity depends not only on the change of binding free energy. According to the aggregation−activation theory, receptors exist in a dynamic equilibrium between active and inactive states.[69] Similarly, enzymes can adopt conformations with different activities. Therefore, the level of inhibition of RT depends on the extent of conformational changes induced by the NNRTIs. Currently, there is no general methodology that would allow correlating the receptor conformation with its biological activity. Because there are many chemical groups of NNRTIs that can adopt several binding geometries, finding a general method for predicting their activity would be problematic. However, for a group of closely related compounds that (presumably) have the same binding mode, it can be assumed that there are no major differences between the conformations of RT in complexes with those compounds. It is therefore expected that for structurally similar compounds direct correlation between binding free energy and biological activity can be found, provided there are no other factors that can interfere.

In order to obtain more information, we extended our study on three additional sets of azole NNRTIs: tetrazoles (set 4), 1,2,3-thiadiazoles (set 5), and imidazoles (set 6). Although those compounds have a different core heterocycle, their structure is closely related to the triazoles from sets 2 and 3. Glide (SP and XP) was selected for docking because this program showed relatively good performance in this study. The results are presented in Tables 9 and 10. In set 4, both of the docking protocols gave poor correlations, except for Glide SP in 2RKI, which achieves good correlation for *evdw* and *energy* for valid poses. An important note is that the activity values for set 4 are given as IC_{50}, while for all other sets as EC_{50}. In set 5, docking to 2RKI was not very successful both in terms of pose prediction and activity prediction. Only *gscore* for valid poses (SP) shows some statistically significant negative correlation. The results for 3DLG are significantly better—the returned poses were more consistent and accurate, which supports the assumption that thiadiazoles and benzophenones bind with RT in a similar manner. Most of the SP scores show low positive correlation with the biological activity. The results for the XP mode show better correlation, which is highest for the *emodel*, *evdw*, and *energy* values. Surprisingly good results were obtained by both docking protocols in set 6, reaching r_s values well over 0.900 in some cases. Similar to the results for set 1, the best correlated terms are *gscore*, *emodel*, *evdw*, and *energy*.

These results indicate that the docking programs accuracy is highly dependent on the choice of a ligand set. Still, it is not clear what causes the differences in performance for different sets. Even though the scoring functions of docking tools use simplified intra- and inter-molecular interaction terms, this cannot explain inconsistent scoring of ligands bearing high structural resemblance. For example, if desolvatation energy or van der Waals interactions play crucial role in the binding of compounds in set 6, it is very unlikely that it has only minor role in set 2.

Critical for any comparative statistical analysis is the quality of target data. Considering the fact that activity data for set 2 and 3 come from the same publication and that the values are given only as means without standard deviations, plus that the measurements were cell-based assays, there is also a possibility that the target data was not accurate. In that case, the only way to validate the data would be repeating the biological activity measurements in at least two independent laboratories. If we assume that the data for sets 2 and 3 are inaccurate and only small to moderate errors are present, we can get some useful information by recalculating the correlation coefficients after aggregating the nearest samples. The biological activity range (in pEC_{50}) was divided into intervals in which all compounds were assigned an equal rank value. Intervals with width ranging from 0.125 to 0.750 were used. Spearman's ρ coefficients were calculated using modified eq 5 for tied ranks formed by aggregation. Like before, Glide SP and XP were used for this analysis. Because for set 1 and 6 the highest positive correlations were obtained for *gscore*, *emodel*, *evdw*, and *energy*, these values were averaged and plotted against aggregation interval width (Figures 4−7).

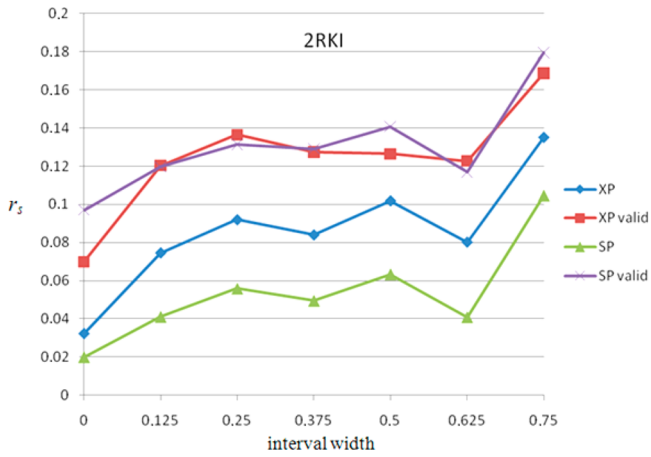


Figure 4. Average r_s for *gscore*, *emodel*, *evdw*, and *energy* for set 2.

In set 2 for both 2RKI and 3DLG, a slow increase in the mean r_s values can be observed. In set 3 for 2RKI, there is only

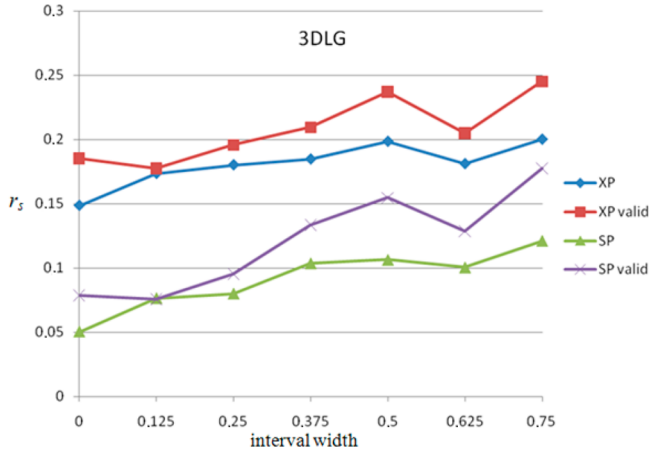


Figure 5. Average r_s for *gscore*, *emodel*, *evdw*, and *energy* for set 2.

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

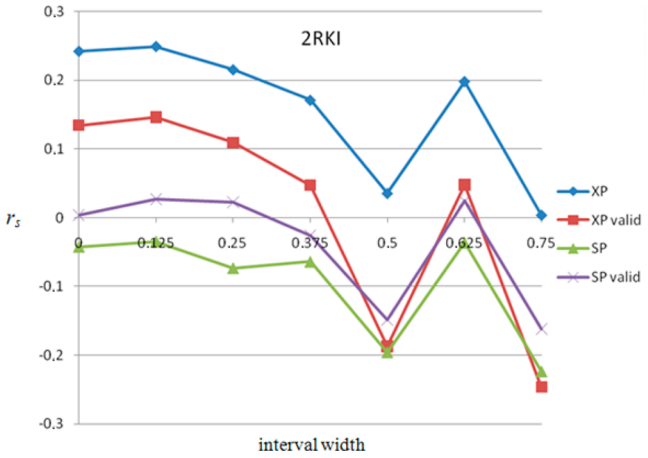


Figure 6. Average r_s for *gscore*, *emodel*, *evdw*, and *energy* for set 3.

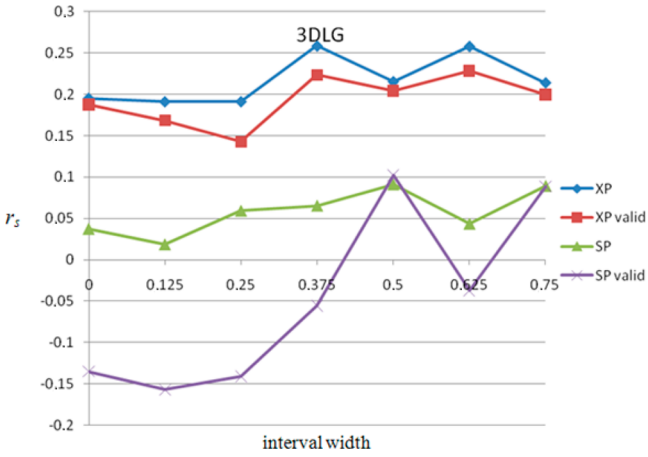


Figure 7. Average r_s for *gscore*, *emodel*, *evdw*, and *energy* for set 3.

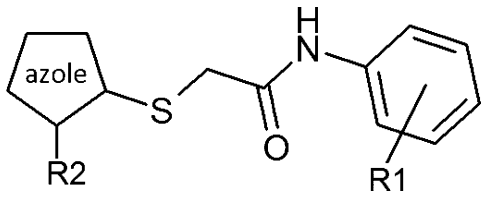
a slight increase for the smallest interval of 0.125. As the interval is incremented, the r_s value drops down below zero, resulting in a negative correlation. For 3DLG, after an initial drop, the mean r_s values start to rise similarly to set 2. It should be noted that with the rising number of tied ranks, the T_X parameter in eq 5 increases, resulting in lower r_s values. For this reason, even if correlation between aggregated sample groups increases, it must be corrected for the uncertainty resulting from aggregation; therefore, r_s will never reach the maximum value of 1 or −1.

As expected, aggregation caused the increase in the r_s values for set 2. This indicates two possible situations: (1) The target data are accurate, but scoring functions can only distinguish between ligands with large differences in activity. (2) The scoring functions are accurate, but the activity data are not. It can also be a combination of those two situations. For set 3, the results are less clear because the activity range in this set is only 2 log units, plus it contains less compounds. This is the reason for negative r_s values for larger aggregation intervals.

To further pursue the possibility of inaccurate biological activity data in sets 2 and 3, the structure−activity relationship for several groups of compounds was explored. Assuming that a given chemical group has the same qualitative effect on the activity of similar ligands, the substituents effects can be compared between different groups of azole NNRTIs (Table 11).

As shown in Table 11, there are some differences between set 2 and the other groups of inhibitors. The most important observation is that in set 2 bromine in position 2 of the phenyl ring is reported to increase the activity relatively to chlorine, but in four out of five other sets, the 2-Cl substituted compounds are more active than 2-Br or 2-Br,4-CH_3 analogs. In one group of 1,2,3-thiadiazoles, the relative activity of chloro and bromo analogs is dependent on the substitution in the second phenyl ring. Although it cannot be excluded that substitution in the remote part of a molecule changes the effect on activity of substitution in the other part of the molecule, there is a strong indication that the chlorine substitution in position 2 of aniline is optimal. Larger halogens like bromine and iodine decrease the activity of this group of NNRTIs. Given the fact that activity effects for other substituents are generally in good agreement, it suggests that the data for set 2 may indeed contain some inaccuracies. For example, the compound with the symbol 7h from set 2 and compound 4g discussed by Wang et al.[50] have the same chemical structure. While the reported EC_{50} values for the reference compound efavirenz are in good agreement: 0.0004 (De La Rosa et al.[31]) versus 0.001 (Wang et al.), the activities for the compound 7h/4g are given as 0.0004 (De La Rosa et al.) and >10 (Wang et al.), which is at least 4 orders of magnitude different. In fact, every measurement is

Table 11. SAR of selected azole-based NNRTIs



azole type	effect on activity
1,2,4-triazole (set 2)[31]	**R1:** 2-Br = 2-Br,4-CH_3 > 2-I > 2-Cl > 2,4-diCl > 2-CF_3> 2,3-diCH_3 > 2,5-diCl > 4-Cl
	R2: p-tolyl >1-naphtyl
tetrazole (set 4)[32]	**R1:** 2-Br > 2-I
1,2,3-thiadiazole (set 5)[37]	**R1:** 2-Cl > 2-Br > 2-Br,4-CH_3> 2,3-diCH_3 > 4-Cl
1,2,3-thiadiazole (set 5)[38]	**R1:** 2-Br, 4-CH_3 > 2-Br ≥ 2-Cl or 2-Cl > 2-Br > 2-Br,4-CH_3
imidazole (set 6)[39]	**R1:** 2-Cl > 2-Br > 2-Br,4-CH_3
	R2: 1-naphtyl > p-tolyl
thiazole[70]	**R1:** 2-Cl > 2-Br
1,2,4-triazole[50]	**R1:** 2-Cl > 2,4-diCl > 2,5-diCl > 2-CF_3 > 2,3-diCH_3 > 4-Cl > 2-Br

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

associated with errors, and the biological activity measurements are especially susceptible as there are many potential error sources. It is not possible for any of the experimental data used in this study to be error free, but the number and magnitude of inaccuracies remain unknown. Therefore, even if some indications were found, no definite conclusion can be made about the source of poor correlation in sets 2 and 3.

Overview of the Relative Biological Activity Prediction Performance. Predominately, for comparisons and applications of molecular docking, only main scoring functions are used. Such functions are constructed and parametrized in a manner to yield appreciable results in most cases of ligand−protein complexes. The correlation coefficients for the main scores of all docking protocols and functions used in this study are plotted for combined sets 1, 2, and 3 (Figure 8) and set 1, where the scoring performances were mostly the best (Figure 9).

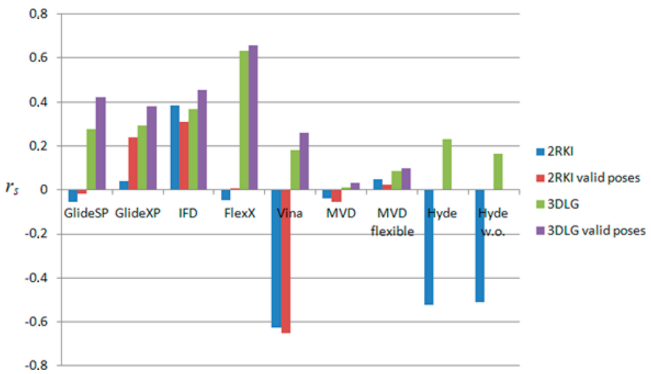


Figure 8. Correlation coefficients for main scores of docking protocols and Hyde for combined sets 1, 2, and 3. Hyde w.o.: Hyde score without optimization.

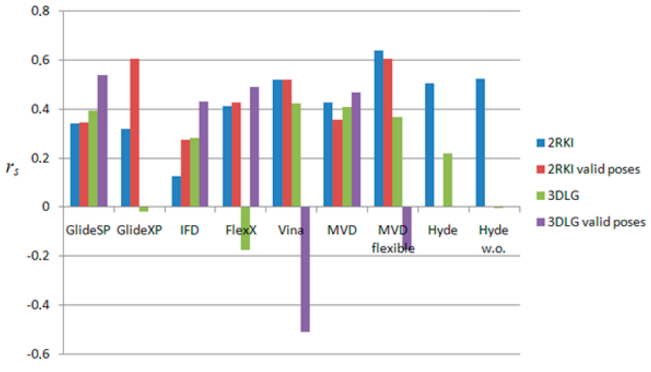


Figure 9. Correlation coefficients for main scores of docking protocols and Hyde for set 1. Hyde w.o.: Hyde score without optimization.

For lead identification purposes, a scoring function should be able to discern between compounds with significantly different binding modes and biological activities. Even though most of the test compounds were in sets 2 and 3 where all scoring functions showed low or no correlation for relative activity predictions, most of the functions show some correlation for combined sets (Figure 8). This results from the aforementioned ability to correctly rank compounds with large differences of the biological activity, which covered 6 logarithmic units. The highest absolute r_s values were achieved by FlexX for 3DLG and Vina for 2RKI. The latter, however, had negative correlation resulting from overrating the ligands tightly fitting the receptor. This precludes the use of Vina for virtual screening of diverse ligand sets that have various binding modes. Given the rather good binding mode prediction, Vina can be used for poses generation, which should be subsequently scored with some other program to reject the false positives and correctly rank the remaining ligands. Similar results to Vina were obtained for Hyde, but because it is not a docking tool, its high dependency on the initial ligand pose and receptor conformation limits its use for lead identification. Most of the programs performed better for 3DLG, probably because this receptor conformation has larger binding site volume, which facilitated docking of larger ligands. The results for 2RKI and 3DLG were closest for IFD due to reduction of conformational differences between those two receptors.

Figure 9 shows correlation coefficients obtained for set 1. This represents the situation analogous to lead optimization, where only small chemical modifications are introduced to the lead compound in order to increase its biological activity. A computational method employed for predicting how small modifications will affect biological activity needs to be accurate and reliable. Most of the used scoring functions give significant positive correlations for valid poses, with the exceptions of Vina and MVD with "flexible residues" for 3DLG. Because the MolDock Score [Grid] function used for MVD gives positive correlation for ligands docked to a rigid 3DLG receptor, it indicates that the function is highly sensitive to errors in the binding mode prediction. Even those protocols that gave the highest correlations are not sufficiently accurate to make a confident classification of compounds by their activity. They can, however, be used to reduce the number of lead structure modifications to be synthesized by removing the compounds with the lowest scores. Also, they are suitable for construction of a consensus scoring method or to be used as a part of a regression model.

In this study, we also assessed some of the components of scoring functions and alternative scoring methods, if present. We have found that many of them performed better than main scoring functions. The binding site of RT is formed by hydrophobic amino acid residues. Polar and electrostatic interactions play minor roles in the enzyme inhibition by NNRTIs. Therefore, hydrophobic and van der Waals interactions terms also correlate with the biological activity. For Glide-based docking protocols, those interactions terms are *evdw* and *energy*. The *energy* is a modified Coulomb and van der Waals interactions term constructed in a manner to avoid overly rewarding charges interactions at the expense of dipole−dipole and charge−dipole interactions. The Glide user manual suggests combining the *gscore* and *energy* to increase the enrichment factor in database screenings. It should be noted that the electrostatic energy term *ecoul* in many cases also gave significant but inconsistent correlation (positive in some cases, negative in other). Another Glide scoring function is *emodel*, which is an extension of *gscore* with some additional energy terms. We found that *emodel* performs similarly or better than *gscore*. In other programs, the lipophilic interactions terms, like *lipo* and *ligand lipo* in FlexX and Hyde, respectively, show similar or better r_s values than main scores in set 1. For MVD, where electrostatic interactions were reduced to zero due to simplified charges scheme, the main component of the final score was the van der Waals energy. Introduction of OPLS partial charges did not significantly affect the docking and scoring output (Tables S12−S15, Supporting Information). The default *rerank* function of MVD, which uses several

dx.doi.org/10.1021/ci400427a | J. Chem. Inf. Model. 2013, 53, 3326−3342

additional interaction terms and penalties, generally did not improve the relative activity prediction. For use with the triazoles-HIV RT system, the *rerank* function should be appropriately parametrized (which can be easily done by the user). One needs to be cautious, however, not to cause data overfitting with the modified scoring function.

A general remark from those results is that when designing novel azole-based (and probably most of the other types) NNRTIs, one of the main targets should be maximizing the hydrophobic and van der Waals interactions. Even though the biological activity predictions for sets 2, 3, and 4 were poor, it does not completely exclude the use of docking tools for that purpose. If preliminary studies like this one show only low correlation for a given group of compounds, one can still make a rough activity estimation based on the most important interactions terms. For example, *evdw* and *energy* scores of Glide SP (valid poses, 2RKI) shows no statistically significant correlation, but almost all compounds with pEC_{50} less than 8 are placed in the lower half of the ranking (Figures S7 and S8, Supporting Information). Also, many of the most active compounds are ranked very low, which shows that those scores have greater tendency for false negatives than for false positives. A reduced rate of false positive hits is desired for virtual screening purposes, even at the expense of lower true positive hits rate.

CONCLUSIONS

The main purpose of the presented studies was the evaluation of different molecular docking protocols for their ability to predict the relative biological activity of triazole-based NNRTIs. Several programs, search algorithms, and scoring functions were tested. All of the programs performed similarly, but AutoDock Vina, Molegro Virtual Docker, and Hyde favored shape matching of ligand and binding sites, which limits their use for activity prediction only to compounds closely related to a native ligand of a given receptor conformation.

It was found that predictions of relative biological activity were highly dependent on the ligand set. While all programs were able to make reasonable predictions in set 1, none were able to correctly rank ligands in sets 2 and 3. Because most of the used compounds have very similar structures, it is expected that their binding modes with the HIV-1 RT are also very similar. Therefore, a given docking protocol should perform comparably in most of the test sets. Moreover, NNRTIs interactions with the enzyme are mostly nonspecific, so even the simplest scoring functions should be able to make some useful predictions. Some indications were found that low correlation for sets 2 and 3 may result from the poor quality of target data; however, no definitive conclusion can be made. Biological activity measurements are susceptible to large errors, especially when using cell-based assays; therefore, it is difficult to obtain good quality data. Using rank correlation methods like Spearman's ρ can compensate to some extent for inaccuracies of the target data. It is also possible that the activity in sets 2 and 3 was affected by some specific or nonspecific interactions with the biological system components. In any case, this study shows that choosing the target data is more important than the computational methodology.

None of the scoring methods achieved a perfect ranking (r_s = 1) of ligands by their activities. The highest correlations observed in this study were around 0.7 (Glide XP) for set 1, excluding nontriazole compounds from set 6, where the correlation coefficients were even higher. Spearman's ρ coefficient of 0.7 corresponds to around 0.75 probability of identifying the more active compound from two compounds. Such probability is sufficient to choose a smaller subset of compounds for synthesis from a larger ensemble of similar compounds. For lead optimization, the error chance of 0.25 may be still too high, but using consensus scoring can increase the confidence of prediction, provided that the different scoring methods are not correlated with each other.

The accurate prediction of biological activity is complex and still unsolved problem. Scoring functions of docking programs offer only a greatly simplified and approximated solution; therefore, one should not solely rely on raw docking scores. When docking results are combined with experienced researcher's assessment, they can provide useful information and guide the researcher toward the design of potent compounds. Even though this study is far from extensive, it suggests that different molecular docking methods are able to distinguish higher from lower active azole-based NNRTIs. It should be noted that the task for the docking programs was particularly difficult because all of the used ligands had very similar chemical structures and they were docked not to their respective crystallographic structure of the enzyme. Additionally, all ligands were active inhibitors; unlike typical database screening where most of the compounds are inactive, discriminating two ligands with similar activity is more difficult than discerning an active ligand from an inactive one.

Case studies like this one are very important and should be always performed (if possible) to choose the best methodology for a given ligand−protein system. Benchmark studies of docking methods are often performed for a diverse set of protein targets using default settings of the docking programs. This can give a false image of the docking tools accuracy because such studies are basically benchmarks of the default settings of the programs and not their accuracy itself. The methodology we used for each docking method was not optimized but was based on our general experience with docking programs and HIV-1 reverse transcriptase. The results can be further improved by performing a more systematic search for the optimal docking parameters. Another possibility to obtain more accurate prediction is to employ consensus scoring or by building a regression model based on docking results. These studies are now in progress.

METHODS

Receptor Preparation. Proteins were prepared using the Protein Preparation Wizard in Schrödinger's Maestro suite. All water molecules and hetero groups (except for the ligand) were removed. Preprocessing was carried out with "Assign bond orders", "Add hydrogens", "Create disulfide bonds", "Fill in missing side chains using Prime", and "Fill in missing loops using Prime" options selected. For H-bond assignment in *Refine* step, "Exhaustive sampling" and "Sample at pH Neutral" were selected. Final structure minimization was performed with default settings using the OPLS-2005 force field. Resulting protein structures were used for docking in Glide. For MVD, FlexX, and AutoDock Vina, data was exported to .pdb or .mol2 formats.

Ligand Preparation. Ligands for the use with Glide were prepared with the LigPrep utility. A generation of tautomers and possible protonation states (using Epik, pH 7 ± 2) options were selected, but no alternative states were generated. OPLS-2005 partial charges were assigned to all of the atoms, and structures were optimized. The resulting structures were

exported to the .mol2 format and were used as a starting point for the preparation of ligands for the remaining docking programs. For docking in Molegro, the ligands were imported and prepared using built-in dialogue. The "Assign bonds", "Assign bond orders and hybridization" and "Detect flexible torsions in ligands" options were selected as "Always", "Create explicit hydrogens" as "Never", "Assign Tripos atom types" as "If missing", and "Assign charges (calculated by MVD)" was selected as "Always" for the MVD charges or as "Never" to keep the OPLS-2005 charges. For AutoDock Vina and FlexX internal MVD, Gasteiger and internal canonical charges were assigned, respectively.

Docking. *Glide.* Glide requires prepared receptor grids for docking. The receptor grids for both 2RKI and 3DLG were prepared in the same manner: Grid size was assigned automatically and centered at the centroid of the reference ligand from the crystallographic structure. van der Waals radii of atoms with the absolute value of partial charge less than 0.2 were scaled by 0.9. Hydroxyls of Tyr181, Tyr188, and Tyr318 (chain A) were allowed to rotate. Docking in standard (SP) and extra precision (XP) modes were performed with default settings. Three best poses for every ligand were kept. For the ranking of resultant poses, several scores were used. *GlideScore* (*gscore*), the main Glide SP scoring function based on modified ChemScore function,[41,71] is given by the general formula

$$gscore = 0.065 \times evdw + 0.130 \times ecoul + lipo + hbond + metal + buryp + rotb + site \quad (7)$$

Two components of the function were separately used for ranking the poses: *evdw*, van der Waals energy, and *ecoul*, Coulomb energy. The remaining terms: *lipo*, hydrophobic interactions; *hbonb*, hydrogen bonding; *metal* , metal binding; *buryp*, penalty for buried polar groups; *rotb*, penalty for freezing rotatable bonds; and *site*, polar interactions in the active site were not used because they were not suitable for pose ranking (e.g., equal to 0). For the XP mode, the scoring function has the form

$$XPgscore = evdw + ecoul + ebind + epenalty \quad (8)$$

Similar to the SP mode, only *evdw* and *ecoul* terms were used for the pose ranking. The *ebind* and *epenalty* comprise several terms for hydrophobic interactions, various types of hydrogen bonding, desolvatation energy, and ligand strain penalties and are described in ref 64. Additionally, we used the following scores for pose ranking: *emodel*, energy model that combines the *GlideScore*, nonbonded interaction energy, and internal strain energy; *energy*, modified Coulomb−van der Waals interaction energy; and *eint*, internal torsional energy of the ligand.

Induced Fit Docking. The prepared enzyme structures used for SP and XP docking were used for receptor definition. The receptor grid was centered at a reference ligand from the crystallographic structure, and the size of the grid was selected automatically. The initial docking step was performed using default parameters (receptor and ligand van der Waals scaling = 0.5), and then residues within 6 Å radius of the docked ligand were optimized with Prime. Finally, best poses were redocked to their corresponding receptor conformation using Glide XP (default settings). For pose ranking, the same scores were used as for normal Glide docking, plus additional *IFDScore* which by default equals to *gscore* + 0.05 × Prime energy from the protein refinement calculations.

Molegro Virtual Docker. MVD can perform docking to a rigid receptor or to a receptor with some of the residues defined as flexible. We have found that defining too many (five or more) amino acid side chains as flexible yielded no results; therefore, the limited conformational adaptability of the binding site was simulated by using reduced interactions for selected residues (residue numbers: 100, 103, 106, 188, and 234 for 2RKI and 100, 103, 106, 181, 188, and 234 for 3DLG). In "flexible" residues definition dialogue, the interaction tolerance was increased to 1.1 with strength scaled by 0.9. In the Docking Wizard dialogue, the search sphere (radius 15 Å) was centered manually at reference ligand. *MolDock Score [Grid]* was selected for pose scoring. Grid resolution was set to 0.20 Å, and the "Internal ES" and "Energy Minimization" options were toggled on. The remaining parameters were kept default. For pose clustering, the "Return multiple poses for each run" option was selected. The maximum number of returned poses was set to five, from which three best poses were automatically chosen in *Pose Organizer* for subsequent manual inspection. The poses were ranked using *MolDock Score* and *Rerank Score*. The *MolDock Score* function[43] consist of two terms: E_{inter}, the ligand−protein interaction energy that comprises electrostatic interactions and piecewise linear potential approximation of the steric and hydrogen bonding interactions between all heavy atoms in the ligand and the protein (and waters or cofactors if present), and E_{intra}, the sum of the interactions between ligand atoms, the torsional energy term, and the penalty term for unfeasible ligand conformations. The *Rerank* function is a weighted form of *MolDock Score* with additional terms (25 in total) including the Lennard−Jones approximation of steric interactions (LJ12−6). The default parameters for the *Rerank* score were used.

FlexX and Hyde. FlexX and Hyde programs were used as implemented in the LeadIT software package.[72] The receptor for FlexX was prepared using the Prepare Receptor dialogue. Both protein chains were selected, and the active site was defined to include all residues within 20.0 Å radius of the native ligand. The rest of the preparation steps were carried out automatically. Ligands were imported directly from the .mol2 file with the "Protonate as in aqueous solution" option disabled. Both hard (Maximum Allowed Overlap Volume = 2.9 Å^3) and soft docking (Maximum Allowed Overlap Volume = 100 Å^3) were performed (only the results for the latter were used). For ligand base placement, the hybrid Enthalpy and Entropy strategy was used. The Clash Factor was set to 0.6. Other parameters were kept default. The first three top ranked docking poses were saved for each docking run. The FlexX scoring function[29,44] is given by the general formula

$$score = match + lipo + ambig + clash + rot \quad (9)$$

The poses were ranked basing on *score*, main docking score; *match*, component of main score for matched interacting groups (hydrogen, metal, and specific aromatic interactions), and *lipo*, lipophilic interactions term (for nonpolar atom pairs). The remaining terms, *ambig*, lipophilic interactions (for polar−nonpolar atom pairs); *clash*, penalty for overlapping atoms of a ligand and a protein, and *rot*, weighted number of rotatable bonds, were not used basing on our preliminary study.

The input ligand−protein complexes required for Hyde assessment were generated by Glide SP because this docking method had the highest valid pose prediction rate. Only the valid poses (exported to .mol2 file) from Glide SP docking were used. Scoring was performed from command line using

the hydescorer.exe program. The default settings were used with the exception of the "--optimize" option, which was set to "--optimize 2" to perform the numerical ligand−protein optimization or to "--optimize 0" to skip this step. The results were saved to a .csv file. The Hyde score is composed from two main terms: ΔG_{desolv}, desolvatation energy; and $\Delta G_{H\text{-}bond}$, hydrogen bonding energy.[46] The scoring program returns over 20 different scores and components of main function. In a preliminary study, most of them were tested, and the best four were selected for the ranking of poses: main score, ΔG *desolv*, desolvatation energy; *ligand lipo*, ligand lipophilic contact; and *ligand E desolv*, ligand energy of desolvatation.

Autodock Vina. The receptor proteins and ligands were imported to AutoDock Tools graphic interface and saved in a .pdbqt format. The search space of 24 Å × 22 Å × 20 Å was used for both structures of the HIV-1 reverse transcriptase with the box center (in Å) at x = 9.194, y = 12.417, and z = 16.944 for 2RKI and x = −3.5, y = −33.833, and z = 25.472 for 3DLG. The highest precision (exhaustiveness level 8) was used. The scoring function of Vina is given by eq 10.[45] Only the main score was used for the pose ranking.

$$\text{score} = \Delta G_{\text{attraction}} + \Delta G_{\text{repulsion}} + \Delta G_{\text{H-bond}} + \Delta G_{\text{hydrophobic}} + \Delta G_{\text{torsions}} \quad (10)$$

■ ASSOCIATED CONTENT

Supporting Information

Complete statistical results, structures, and biological activity values of used compounds, comparison between RT crystallographic structures, ligand pose examples, and selected plots of docking scores vs biological activity. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: paneth@p.lodz.pl. Phone: +48 42 631 31 99. Fax: +48 42 631 31 99.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The reported studies are supported by Grant 2011/02/A/ST4/00246 (2012−2017) from the Polish National Center for Science (NCN).

■ REFERENCES

(1) Baldwin, J. J.; Ponticello, G. S.; Anderson, P. S.; Christy, M. E.; Murcko, M. A.; Randall, W. C.; Schwam, H.; Sugrue, M. F.; Springer, J. P.; Gautheron, P.; Grove, J.; Mallorga, P.; Viader, M. P.; McKeever, B. M.; Navia, M. A. Thienothiopyran-2-sulfonamides: Novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. *J. Med. Chem.* **1989**, *32*, 2510−2513.

(2) Von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418−423.

(3) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **2004**, *47*, 1879−1881.

(4) Rahuel, J.; Rasetti, V.; Maibaum, J.; Rüeger, H.; Göschke, R.; Cohen, N. C.; Stutz, S.; Cumin, F.; Fuhrer, W.; Wood, J. M.; Grütter, M. G. Structure-based drug design: The discovery of novel nonpeptide orally active inhibitors of human renin. *Chem. Biol.* **2000**, *7*, 493−504.

(5) Burkhard, P.; Hommel, U.; Sanner, M.; Walkinshaw, M. D. The discovery of steroids and other novel FKBP inhibitors using a molecular docking program. *J. Mol. Biol.* **1999**, *287*, 853−858.

(6) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401−408.

(7) Böhm, H. J.; Banner, D. W.; Weber, L. Combinatorial docking and combinatorial chemistry: Design of potent non-peptide thrombin inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 51−56.

(8) Bressi, J. C.; Verlinde, C. L. M. J.; Aronov, A. M.; Shaw, M. L.; Shin, S. S.; Nguyen, L. N.; Suresh, S.; Buckner, F. S.; Van Voorhis, W. C.; Kuntz, I. D.; Hol, W. G. J.; Gelb, M. H. Adenosine analogues as selective inhibitors of glyceraldehyde-3-phosphate dehydrogenase of trypanosomatidae via structure-based drug design. *J. Med. Chem.* **2001**, *44*, 2080−2093.

(9) Iwata, Y.; Arisawa, M.; Hamada, R.; Kita, Y.; Mizutani, M. Y.; Tomioka, N.; Itai, A.; Miyamoto, S. Discovery of Novel Aldose Reductase Inhibitors Using a protein structure-based approach: 3D-database search followed by design and synthesis. *J. Med. Chem.* **2001**, *44*, 1718−1728.

(10) Kawada, H.; Ebiike, H.; Tsukazaki, M.; Nakamura, M.; Morikami, K.; Yoshinari, K.; Yoshida, M.; Ogawa, K.; Shimma, N.; Tsukuda, T.; Ohwada, J. Lead optimization of a dihydropyrrolopyrimidine inhibitor against phosphoinositide 3-kinase (PI3K) to improve the phenol glucuronic acid conjugation. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 673−678.

(11) Laurini, E.; Da Col, V.; Wünsch, B.; Pricl, S. Analysis of the molecular interactions of the potent analgesic S1RA with the r_1 receptor. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 2868−2871.

(12) Sengupta, S.; Obiorah, I.; Maximov, P. Y.; Curpan, R.; Jordan, V. C. Molecular mechanism of action of bisphenol and bisphenol A mediated by oestrogenreceptor alpha in growth and apoptosis of breast cancer cells. *Br. J. Pharmacol.* **2013**, *169*, 167−178.

(13) Filikov, A. V.; Mohan, V.; Vickers, T. A.; Griffey, R. H.; DanCook, P.; Abagyan, R. A.; James, T. L. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 593−610.

(14) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(15) Caporuscio, F.; Rastelli, G.; Imbriano, C.; Del Rio, A. Structure-based design of potent aromatase inhibitors by high-throughput docking. *J. Med. Chem.* **2011**, *54*, 4006−4017.

(16) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; 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.

(17) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(18) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(19) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDB bind refined set of 800 protein−ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(20) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.

(21) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.

(22) Houston, D. R.; Walkinshaw, M. D. Consensus docking: Improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* **2013**, *53*, 384−390.

(23) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(24) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Modell.* **2002**, *20*, 281−295.
(25) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein−ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.
(26) Vaqué, M.; Ardévol, A.; Bladé, C.; Salvadó, M. J.; Blay, M.; Fernández-Larrea, J.; Arola, L.; Pujadas, G. Protein-ligand docking: A review of recent advances and future perspectives. *Curr. Pharm. Anal.* **2008**, *4*, 1−19.
(27) Sivaprakasam, P.; Tosso, P. N.; Doerksen, R. J. Structure−activity relationship and comparative docking studies for cycloguanil analogs as *Pf*DHFR-TS inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1787−1796.
(28) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S. A new concept for multidimensional selection of ligand conformations (multiselect) and multidimensional scoring (multiscore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.
(29) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(30) Kirschberg, T. A.; Balakrishnan, M.; Huang, W.; Hluhanich, R.; Kutty, N.; Liclican, A. C.; McColl, D. J.; Squires, N. H.; Lansdon, E. B. Triazole derivatives as non-nucleoside inhibitors of HIV-1 reverse transcriptase: Structure−activity relationships and crystallographic analysis. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1131−1134.
(31) De La Rosa, M.; Kim, H. W.; Gunic, E.; Jenket, C.; Boyle, U.; Koh, Y.; Korboukh, I.; Allan, M.; Zhang, W.; Chen, H.; Xu, W.; Nilar, S.; Yao, N.; Hamatake, R.; Lang, S. A.; Hong, Z.; Zhang, Z.; Girardet, J. L. Tri-substituted triazoles as potent non-nucleoside inhibitors of the HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4444−4449.
(32) Muraglia, E.; Kinzel, O. D.; Laufer, R.; Miller, M. D.; Moyer, G.; Munshi, V.; Orvieto, F.; Palumbi, M. C.; Pescatore, G.; Rowley, M.; Williams, P. D.; Summa, V. Tetrazolethioacetanilides: Potent non-nucleoside inhibitors of WT HIV reverse transcriptase and its K103N mutant. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2748−2752.
(33) Tucker, T. J.; Saggar, S.; Sisko, J. T.; Tynebor, R. M.; Williams, T. M.; Felock, P. J.; Flynn, J. A.; Lai, M. T.; Liang, Y.; McGaughey, G.; Liu, M.; Miller, M.; Moyer, G.; Munshi, V.; Perlow-Poehnelt, R.; Prasad, S.; Sanchez, R.; Torrent, M.; Vacca, J. P.; Wan, B. L.; Yan, Y. The design and synthesis of diaryl ether second generation HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) with enhanced potency versus key clinical mutations. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2959−2966.
(34) Zhang, Z.; Xu, W.; Koh, Y. H.; Shim, J. H.; Girardet, J. L.; Yeh, L. T.; Hamatake, R. K.; Hong, Z. A novel nonnucleoside analogue that inhibits human immunodeficiency virus type 1 isolates resistant to current nonnucleoside reverse transcriptase inhibitors. *Antimicrob. Agents Chemother.* **2007**, *51*, 429−437.
(35) Su, D. S.; Lim, J. J.; Tinney, E.; Wan, B. L.; Young, M. B.; Anderson, K. D.; Rudd, D.; Munshi, V.; Bahnck, C.; Felock, P. J.; Lu, M.; Lai, M. T.; Touch, S.; Moyer, G.; DiStefano, D. J.; Flynn, J. A.; Liang, Y.; Sanchez, R.; Prasad, S.; Yan, Y.; Perlow-Poehnelt, R.; Torrent, M.; Miller, M.; Vacca, J. P.; Williams, T. M.; Anthony, N. J. Substituted tetrahydroquinolines as potent allosteric inhibitors of reverse transcriptase and its key mutants. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5119−5123.
(36) Romines, K. R.; Freeman, G. A.; Schaller, L. T.; Cowan, J. R.; Gonzales, S. S.; Tidwell, J. H.; Andrews, C. W., III; Stammers, D. K.; Hazen, R. J.; Ferris, R. G.; Short, S. A.; Chan, J. H.; Boone, L. R. Structure−activity relationship studies of novel benzophenones leading to the discovery of a potent, next generation hiv nonnucleoside reverse transcriptase inhibitor. *J. Med. Chem.* **2006**, *49*, 727−739.
(37) Zhan, P.; Liu, X.; Li, Z.; Fang, Z.; Li, Z.; Wang, D.; Pannecouque, C.; De Clercq, E. Novel 1,2,3-thiadiazole derivatives as HIV-1 NNRTIs with improved potency: Synthesis and preliminary SAR studies. *Bioorg. Med. Chem.* **2009**, *17*, 5920−5927.
(38) Zhan, P.; Liu, X.; Cao, Y.; Wang, Y.; Pannecouque, C.; De Clercq, E. 1,2,3-thiadiazole thioacetanilides as a novel class of potent HIV-1 non-nucleoside reverse transcriptase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5368−5371.
(39) Zhan, P.; Liu, X.; Li, Z.; Fang, Z.; Li, Z.; Pannecouque, C.; De Clercq, E. Synthesis and biological evaluation of imidazole thioacetanilides as novel non-nucleoside HIV-1 reverse transcriptase inhibitors. *Bioorg. Med. Chem.* **2009**, *17*, 5775−5781.
(40) *Glide*, version 5.7; Schrödinger, LLC, New York, 2011.
(41) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(42) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(43) Thomsen, R.; Christensen, M. H. MolDock: A new technique for high-accuracy molecular docking. *J. Med. Chem.* **2006**, *49*, 3315−3321.
(44) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(45) Trott, O.; Olson, A. J. AutoDockVina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(46) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an integrated description of hydrogen bonding and dehydration: Reducing false positives in virtual screening using the HYDE scoring function. *Chem. Med. Chem.* **2008**, *3*, 885−897.
(47) Zhan, P.; Li, X.; Li, Z.; Chen, X.; Tian, Y.; Chen, W.; Liu, X.; Pannecouque, C.; De Clercq, E. Structure-based bioisosterism design, synthesis and biological evaluation of novel 1,2,4-triazin-6-ylthioacetamides as potent HIV-1 NNRTIs. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 7155−7162.
(48) Gagnon, A.; Amad, M. H.; Bonneau, P. R.; Coulombe, R.; DeRoy, P. L.; Doyon, L.; Duan, J.; Garneau, M.; Guse, I.; Jakalian, A.; Jolicoeur, E.; Landry, S.; Malenfant, E.; Simoneau, B.; Yoakim, C. Thiotetrazole alkynylacetanilides as potent and bioavailable non-nucleoside inhibitors of the HIV-1 wild type and K103N/Y181C double mutant reverse transcriptases. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4437−4441.
(49) Gagnon, A.; Landry, S.; Coulombe, R.; Jakalian, A.; Guse, I.; Thavonekham, B.; Bonneau, P. R.; Yoakim, C.; Simoneau, B. Investigation on the role of the tetrazole in the binding of thiotetrazolylacetanilides with HIV-1 wild type and K103N/Y181C double mutant reverse transcriptases. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 1199−1205.
(50) Wang, Z.; Wu, B.; Kuhen, K. L.; Bursulaya, B.; Nguyen, T. N.; Nguyen, D. G.; He, Y. Synthesis and biological evaluations of sulfanyltriazoles as novel HIV-1 non-nucleoside reverse transcriptase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4174−4177.
(51) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: An evaluation based on the statistical analysis of enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511−524.
(52) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50*, 186−193.
(53) Lima, A. N.; Philot, E. A.; Perahia, D.; Braz, A. S. K.; Scott, L. P. B. GANM: A protein−ligand docking approach based on genetic algorithm and normal modes. *Appl. Math. Comput.* **2012**, *219*, 511−520.
(54) Ding, F.; Yin, S.; Dokholyan, N. V. Rapid flexible docking using a stochastic rotamer library of ligands. *J. Chem. Inf. Model.* **2010**, *50*, 1623−1632.

(55) Borrelli, K. W.; Cossins, B.; Guallar, V. Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J. Comput. Chem.* **2010**, *31*, 1224−1235.
(56) Antes, I. DynaDock: A new molecular dynamics-based algorithm for protein−peptide docking including receptor flexibility. *Proteins* **2010**, *78*, 1084−1104.
(57) Whalen, K. L.; Chang, K. M.; Spies, M. A. Hybrid steered molecular dynamics-docking: An efficient solution to the problem of ranking inhibitor affinities against a flexible drug target. *Mol. Inf.* **2011**, *30*, 459−471.
(58) Floquet, N.; Marechal, J. D.; Badet-Denisot, M. A.; Robert, C. H.; Dauchez, M.; Perahia, D. Using normal modes analysis as a prerequisite for drug design: Application to matrix metalloproteinase inhibitors. *FEBS Lett.* **2006**, *580*, 5130−5136.
(59) Cavasotto, C. N.; Orry, A. J. W.; Abagyan, R. A. The challenge of considering receptor flexibility in ligand docking and virtual screening. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 423−440.
(60) Chandrika, B.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discovery Today* **2009**, *14*, 394−400.
(61) Schrödinger Suite 2011 *Induced Fit Docking* protocol; *Glide* version 5.7, Schrödinger, LLC, New York, 2011; *Prime*, version 3.0, Schrödinger, LLC, New York, 2011.
(62) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534−553.
(63) Kendall, M. G. *Rank Correlation Methods*, 4th ed.; Griffin: London, 1970; pp 29−48.
(64) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(65) Jorgensen, W. L.; Tirado-Rives, J. The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(66) *LigPrep*, version 2.5; Schrödinger, LLC, New York, 2011.
(67) Schrödinger Suite 2011 *Protein Preparation Wizard*; *Epik*, version 2.2, Schrödinger, LLC, New York, 2011; *Impact*, version 5.7, Schrödinger, LLC, New York, 2011; *Prime*. version 3.0, Schrödinger, LLC, New York, 2011.
(68) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.
(69) Monod, J.; Wyman, J.; Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **1965**, *12*, 88−118.
(70) Zhan, P.; Wang, L.; Liu, H.; Chen, X.; Li, X.; Jiang, X.; Zhang, Q.; Liu, X.; Pannecouque, C.; Naesens, L.; De Clercq, E.; Liu, A.; Du, G. Arylazolyl(azinyl)thioacetanilide. Part 9: Synthesis and biological investigation of thiazolylthioacetamides derivatives as a novel class of potential antiviral agents. *Arch. Pharm. Res.* **2012**, *35*, 975−986.
(71) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Design* **1997**, *11*, 425−445.
(72) *LeadIT 2.1.0*; BioSolveIT GmbH: St. Augustin, Germany, 2012.