JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 3584−3599

pubs.acs.org/jcim

Getting Docking into Shape Using Negative Image-Based Rescoring

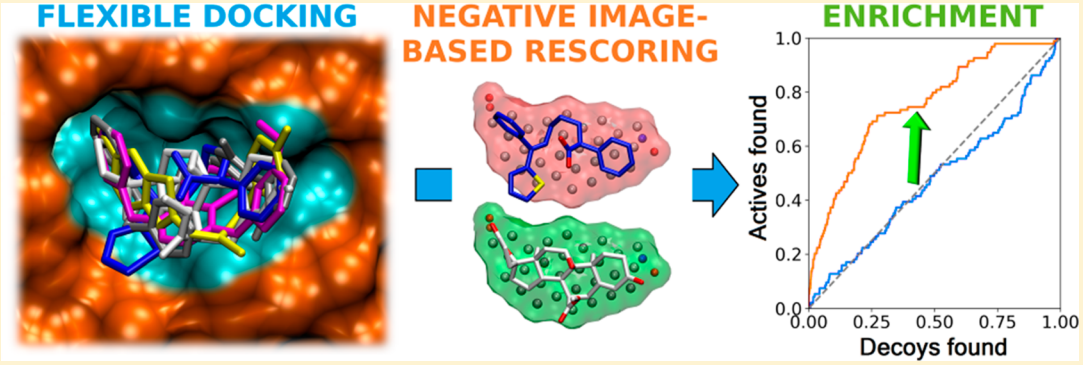
Sami T. Kurkinen,[†] Sakari Lätti,[†] Olli T. Pentikäinen,[†,‡] and Pekka A. Postila*[,§]

[†]Institute of Biomedicine, Kiinamyllynkatu 10, Integrative Physiology and Pharmacy, University of Turku, FI-20520 Turku, Finland

[‡]Aurlide Ltd., FI-21420 Lieto, Finland

[§]Department of Biological and Environmental Science, University of Jyvaskyla, P.O. Box 35, FI-40014 Jyvaskyla, Finland

Supporting Information



ABSTRACT: The failure of default scoring functions to ensure virtual screening enrichment is a persistent problem for the molecular docking algorithms used in structure-based drug discovery. To remedy this problem, elaborate rescoring and postprocessing schemes have been developed with a varying degree of success, specificity, and cost. The negative image-based rescoring (R-NiB) has been shown to improve the flexible docking performance markedly with a variety of drug targets. The yield improvement is achieved by comparing the alternative docking poses against the negative image of the target protein's ligand-binding cavity. In other words, the shape and electrostatics of the binding pocket is directly used in the similarity comparison to rank the explicit docking poses. Here, the PANTHER/ShaEP-based R-NiB methodology is tested with six popular docking softwares, including GLIDE, PLANTS, GOLD, DOCK, AUTODOCK, and AUTODOCK VINA, using five validated benchmark sets. Overall, the results indicate that R-NiB outperforms the default docking scoring consistently and inexpensively, demonstrating that the methodology is ready for wide-scale virtual screening usage.

1. INTRODUCTION

Structure-based drug discovery is increasingly turning toward *in silico* methods such as molecular docking for expediency and cost efficiency.[1−5] Docking aims to predict accurately both the bioactive binding pose and the affinity of a ligand forming the complex with its receptor. Docking sampling, generating alternative ligand binding poses against the receptor's binding site, is performed using incremental construction, matching algorithms, or stochastic methods such as Monte Carlo and genetic algorithms. Docking scoring, which is roughly divided into force-field-based, empirical, or knowledge-based methods, ranks the generated ligand−receptor complexes. Depending on the scoring method, noncovalent bonding interactions and also hydrogen bonding (H-bonding), hydrophobic effect, and even binding entropy can be summed for the final score.[1,2,5,6]

Docking sampling treats either only the ligand flexibly or both the ligand and the receptor adjust reciprocally. As the number of degrees of freedom increases, also the computational costs of docking simulations increase. Although popular these days, it is debatable if either the flexible or induced-fit docking are suitable for high-throughput virtual screening as opposed to much lighter rigid docking simulations.[2,6] A robust metric for assessing sampling is to compare the predicted poses against the experimentally verified poses.[7−11] However, for example, X-ray crystal structures are only snapshots of the dynamic recognition process, and thus, both the ligand and the receptor can have alternative reciprocal conformations.[12,13] Even so, the docking generally samples the "correct" poses excellently or at least reasonably well.[1,5,14,15]

Despite its potential, docking has disadvantages that must be acknowledged. First, the ability of the algorithms to separate active ligands from the inactive ones is highly dependent on the target protein or its conformation.[13,16] Second, even if the right conformer is sampled, it is frequently given too low a score.[14,17,18] Third, the success is dependent also on the software and/or applied settings, and unfortunately, without verification, it is difficult to make the needed adjustments.[9,19,20] Even when relying on benchmarking, the enrichment metrics such as the area under the curve (AUC) might give too rosy a picture because the early enrichment could remain too low for effective drug discovery.[21−23]

Received: May 8, 2019
Published: July 10, 2019

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

As it stands, there exists a plethora of postprocessing techniques, involving force field-based optimization and evaluation steps, for improving the docking results. Due to the high cost of these methods, e.g., SIE (Solvated Interaction Energy) or MM/GB(PB)SA (Molecular Mechanics with Generalized Born or Poisson−Boltzmann and Surface Area solvation) calculations,[24−26] their implementation is limited to the top-ranked compounds in the multistep workflows. Instead, the alternative conformers outputted by the sampling could already contain the bioactive poses, and effective rescoring could rank them correctly. On a case-by-case basis, the docking performance can even be improved by calculating a consensus score that combines the output of several functions.[27−29]

In the negative image-based rescoring (R-NiB; Figure 1),[21] molecular recognition is not considered as the sum of its parts as in standard docking, but the focus is put on the shape/electrostatics complementarity between the ligand and its binding cavity in its entirety. A negative image, encompassing both the key shape and charge features of the cavity (Figure 1), is generated using PANTHER.[22] In the negative image-based (NIB) screening,[22,25,30,31] this cavity-based drug-like NIB model or pseudoligand is used to dock rigidly the *ab initio*-generated ligand conformers via geometry optimization using ShaEP.[32] In contrast, in R-NiB, the shape/charge of the conformers originating from flexible docking is directly compared against the negative image without the optimization (a.k.a. docking). The PANTHER/ShaEP-based rescoring (~2−4 ms/comp.) is much faster to compute than the initial flexible docking (e.g., PLANTS: ~40−80 ms/comp.).[21]

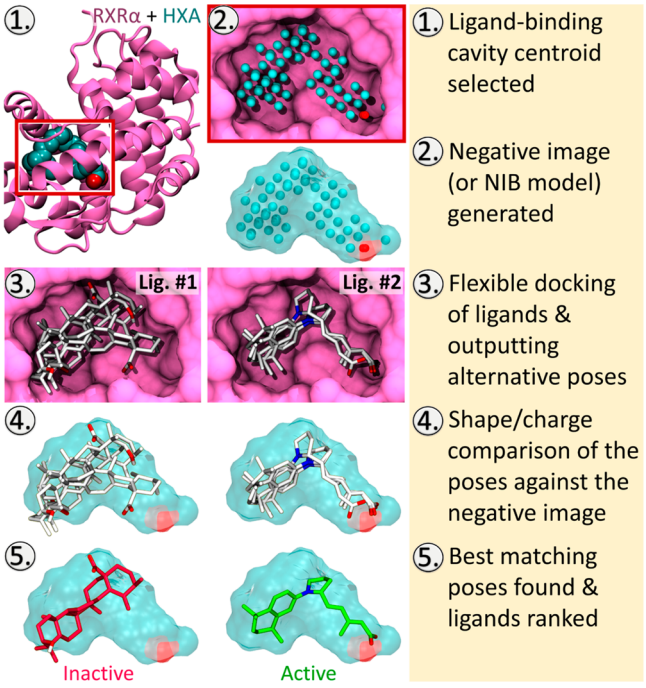


Figure 1. Negative image-based rescoring step-by-step. The negative image-based rescoring (R-NiB) protocol follows five steps: (1) Ligand-binding cavity and its centroid are selected from the protein 3D structure (a cartoon model of RXRα with the bound docosa hexaenoate or HXA; PDB: 1MV9).[33] (2) Negative image or NIB (negative image-based) model (transparent surface), composed of neutral filler atoms (cyan spheres) and negative cavity point (red sphere), is generated using PANTHER.[22] (3) Flexible molecular docking (e.g., VINA[34]) is performed for the ligands (e.g., lig. #1 and lig. #2 or C44184559 and CHEMBL2085503 in the DUD-E set for the RXRα[18]), and several (e.g., N = 3) alternative docking poses (stick models with white backbone) are outputted for each compound. (4) Cavity-based rescoring or the shape/charge comparison of docking poses (one at a time!) is used with the NIB model without geometry optimization using ShaEP.[32] (5) Comparison produces similarity scores (from 1 to 0) for each docked pose, and this information is used to rank the individual docking poses and the ligands. Based on the R-NiB ranking, compounds can be categorized or predicted as inactive (red stick model; e.g., lig. #1) or active (green stick model; e.g., lig #2). Note that the steps involved in the protein or ligand preparation for NIB model generation or docking, visual inspection of the best-ranked poses, or potential benchmarking efforts are omitted for brevity. The figure was created using BODIL[35] and Visual Molecular Dynamics or VMD 1.9.2.[36]

In a prior study,[21] the NIB methodology was modified to facilitate the rescoring of explicit docking solutions of PLANTS.[37] The benchmarking was performed with 11 targets using validated test sets,[38,39] and despite its ultrafast speed and relatively simple premise, R-NiB (Figure 1) was able to improve the flexible docking enrichment with multiple targets. Notably, the yield improvements did not require specific tinkering of the software settings.[21] Here, the aim was to determine if R-NiB (Figure 1) is as efficient with the other widely used docking software as it is with PLANTS. Accordingly, R-NiB was paired in addition to PLANTS with GOLD,[40] GLIDE,[41,42] DOCK,[43] AUTODOCK,[44] and AUTODOCK VINA[34] and tested on five different targets.[39]

While there is software- and target-specific differences, R-NiB (Figure 1) consistently improved the performance of the selected docking algorithms. Overall, R-NiB also produced higher enrichment than the rescoring algorithm SMINA,[45] although the latter method excelled in the very early enrichment with many targets and docking software. Given the clear-cut nature of the results, it is suggested that R-NiB should be routinely paired with flexible docking simulations in virtual screening assays to facilitate efficient drug discovery.

2. RESULTS

2.1. Selecting Targets for Benchmarking. Target proteins (Figure 2; Table 1) were selected based on the dissimilarities in their ligand-binding cavities shape, size, and hydrophobicity or the level of difficulty observed in the prior docking or negative image-based rescoring (R-NiB; Figure 1) efforts.[21] As a result, the composition and the number of filler atoms or charge points in the final negative images or models, reflecting the cavity shape/electrostatics, vary markedly between the targets (N = 44−79; Figure 2). The five target proteins are valid drug discovery targets for which there exist several high-resolution X-ray crystal structures (Table 1; Figure 2) and commonly used benchmarking test sets (Table 1).[39]

The mineralocorticoid receptor (MR; Figure 2A) has a well-defined and mostly hydrophobic ligand-binding site typical for steroid-binding nuclear receptors; i.e., high-affinity binding requires a tight fit between the ligand and the cavity. In contrast, the ligand-binding site of neuraminidase (NEU; Figure 2B) is a small opening on the protein surface; this open-endedness of the target cavity is a mounting challenge for flexible docking algorithms as demonstrated by a case study involving prion protein.[51] Although the retinoid X receptor alpha (RXRα; Figures 1 and 2C) is also a nuclear receptor with a hydrophobic binding site, its pocket is a lot larger than that of MR. The binding site of cyclooxygenase-2 (COX-2; Figure 2D) is a well-defined space with both hydrophobicity- and charge-driven characteristics important for inhibitor binding. Finally, phos-

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

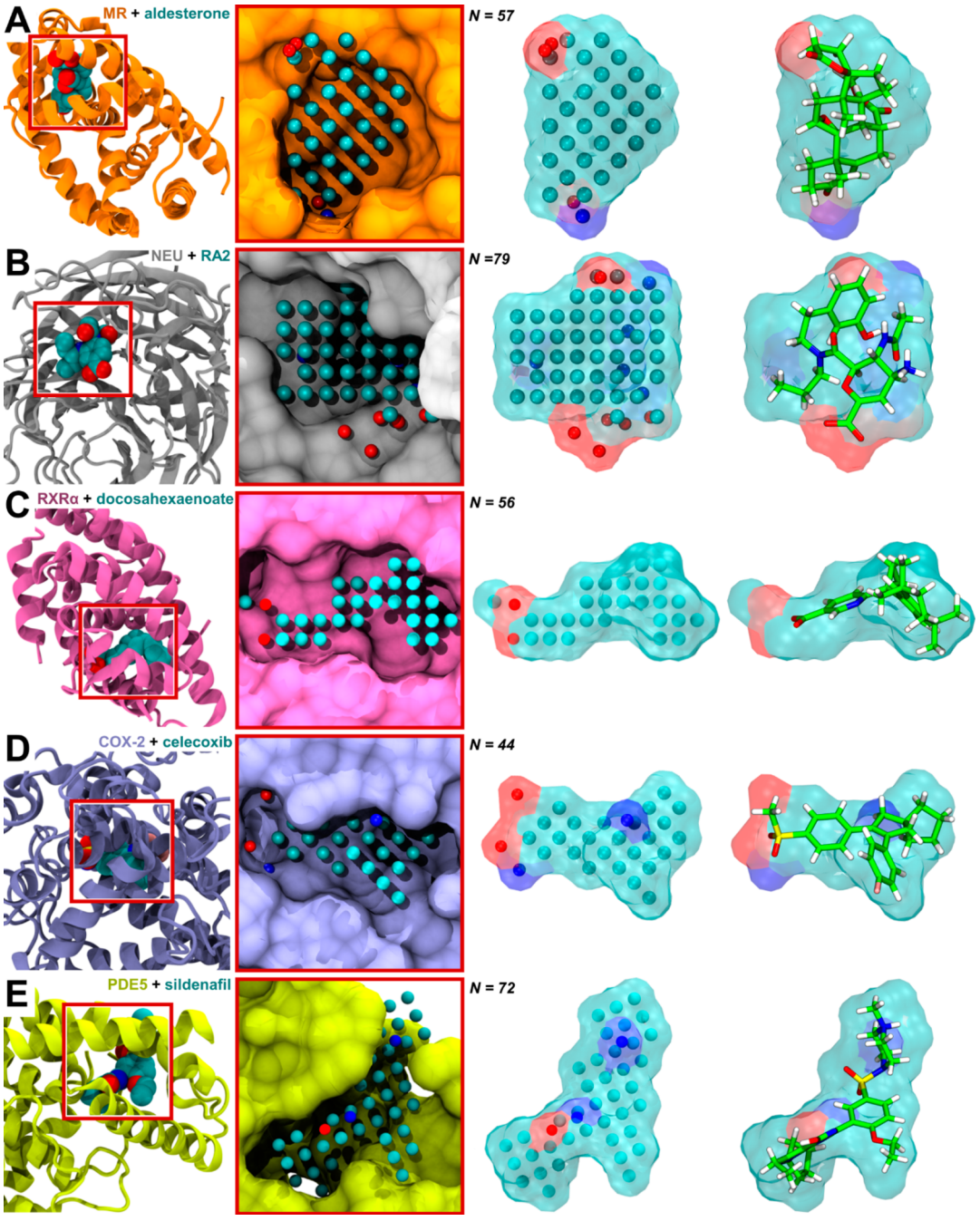


Figure 2. Negative images of target proteins' ligand-binding cavities. On the left, 3D structures of target proteins (cartoon models) with co-crystallized ligands (CPK models) at binding cavities. In the middle, cross sections of binding cavities (opaque surfaces) in close ups (red boxes). PANTHER[22]-generated negative images or NIB (negative image-based) models are composed of neutral filler atoms and negatively or positively charged cavity points (cyan/red/blue spheres). On the right, NIB models are shown with space-filling transparent surfaces either with cavity points or an active ligand from PLANTS[37] docking (sticks with a green backbone). Both the shape and volume (N = 44−79) of NIB models vary substantially between (A) mineralocorticoid receptor (MR; PDB: 2AA2),[49] (B) neuraminidase (NEU; PDB: 1B9V),[50] (C) retinoid X receptor alpha (RXRα; PDB: 1MV9; different angle shown in Figure 1),[33] (D) cyclooxygenase-2 (COX-2; PDB: 3LN1),[46] and (E) phosphodiesterase-5 (PDE5; PDB: 1UDT).[29] NIB models aim to encompass only those cavity sections needed for ligand binding instead of filling the cavities to the brim. The figure was created using BODIL[35] and Visual Molecular Dynamics or VMD 1.9.2.[36]

phodiesterase-5 (PDE5; Figure 2E), whose ligand-binding cavity is spacious and contains plenty of water, was chosen due to the challenge it has presented for flexible docking and negative image-based (NIB) screening as well as R-NiB in prior studies.[21,22,25]

2.2. Default Docking Scoring—Better Than Guessing. The superiority between different docking softwares (Table 1) is under constant debate because the success of molecular docking in separating the active ligands from the inactive compounds is software and target specific.[7−11] Furthermore, the success of the different docking algorithms is difficult to assess reliably via enrichment comparisons if the software in question skips a substantial and variable number of ligands during the docking. This is especially the case with the early enrichments, where a

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Table 1. Benchmark Ligand Sets and Protein 3D Structures

Target protein[a]	RXRα	COX-2	PDE5		MR	NEU
PDB code	1MV9	3LN1	1UDT[b]	1XOZ[b]	2AA2	1B9V
Resolution (Å)	1.9	2.4	2.3	1.37	1.95	2.35
ligs[c]	131	435	398	398	94	98
decs[c]	6935	23,136	27,520	27,520	5146	6197

[a]Retinoid X receptor alpha (RXRα),[33] cyclooxygenase-2 (COX-2),[46] phosphodiesterase-5 (PDE5),[47,48] mineralocorticoid receptor (MR),[49] neuraminidase (NEU).[50] [b]Only PDB entry 1UDT[48] was used for docking. Both 1UDT[48] and 1XOZ[47] were used in NIB (negative image-based) model generation. [c]Number of active ligands (*ligs*) and decoy molecules (*decs*) after performing ligand preparation with LIGPREP in MAESTRO (status before docking screening).

Table 2. Performance of Molecular Docking Algorithms in Benchmarking[a]

			GLIDE			DOCK			
		PLANTS	HTVS	SP	GOLD	Def.	Opt.	VINA	AUTODOCK
RXRα	AUC	0.77 ± 0.02	0.68 ± 0.03	0.83 ± 0.02	0.76 ± 0.02	0.41 ± 0.02	0.52 ± 0.03	0.82 ± 0.02	***0.88 ± 0.02***
	EFd 1%	11.5	44.3	***65.6***	16.8	4.5	8.4	40.5	54.2
	EFd 5%	37.4	50.4	***77.1***	35.1	6.8	13.0	55.7	72.5
	docked *ligs%/decs%*	100/100	51/31	82/67	100/100	26/45	73/69	100/100	100/100
COX-2	AUC	0.66 ± 0.01	0.66 ± 0.01	***0.74 ± 0.01***	0.71 ± 0.01	0.61 ± 0.01	0.64 ± 0.01	***0.76 ± 0.01***	0.61 ± 0.01
	EFd 1%	5.7	24.1	37.2	12.8	11.3	11.3	33.8	3.0
	EFd 5%	21.6	35.2	46.7	38.1	20.2	22.1	***47.4***	11.7
	docked *ligs%/decs%*	100/100	58/40	83/74	100/100	52/35	83/69	100/100	100/100
PDE5	AUC	***0.78 ± 0.01***	0.67 ± 0.01	***0.76 ± 0.01***	***0.76 ± 0.01***	0.50 ± 0.01	0.54 ± 0.01	0.64 ± 0.02	0.61 ± 0.02
	EFd 1%	***11.3***	7.0	10.3	9.0	0.3	1.5	**11.3**	5.8
	EFd 5%	28.1	23.6	32.4	**30.7**	4.8	6.0	22.1	14.1
	docked *ligs%/decs%*	100/100	83/84	94/99	100/100	89/94	96/97	100/100	100/100
MR	AUC	***0.55 ± 0.03***	0.48 ± 0.03	0.50 ± 0.03	0.47 ± 0.03	0.41 ± 0.03	0.42 ± 0.03	0.53 ± 0.03	***0.60 ± 0.03***
	EFd 1%	3.2	7.4	***12.8***	2.1	0.0	0.0	7.4	***12.8***
	EFd 5%	19.1	9.6	19.1	7.4	1.1	0.0	9.6	***24.5***
	docked *ligs%/decs%*	100/100	13/19	34/41	100/100	32/47	60/69	100/100	100/100
NEU	AUC	***0.85 ± 0.01***	0.61 ± 0.03	0.82 ± 0.03	0.69 ± 0.03	***0.83 ± 0.03***	***0.82 ± 0.03***	0.56 ± 0.03	0.62 ± 0.03
	EFd 1%	4.1	***19.4***	18.4	2.0	12.2	14.3	0.0	0.0
	EFd 5%	32.7	31.6	43.9	9.2	34.7	**36.7**	4.1	4.1
	docked *ligs%/decs%*	100/100	96/97	100/100	100/100	99/98	100/99	100/100	100/100

[a]Best docking results are bolded and in italics, if the values are within the error margin. Only those AUC values that are within the error margin are highlighted.

difference of only a few docked ligands can skew the results to either direction.

Here, in order to compare the docking results of different software reliably, the skipped molecules were added to the bottom of the ranking list at even ratios that correspond to random picking. Flexible docking, especially regarding scoring, was also done using the default settings of each software to limit the amount of computations to a manageable level and make the comparison unbiased. In practice, most of the tested algorithms (e.g., DOCK) have multiple and potentially better-suited scoring functions for the targets than default scoring. As per expectation, the performance of the docking software in benchmarking varied a great deal. Usually, default docking scoring was a lot or at least slightly better than a proverbial coin toss in recognizing the active ligands from the inactive decoys (Table 2; Figure 3; Figures S1−S10).

Depending on the test set, PLANTS docking produced AUC values within the range from 0.55 ± 0.03 to 0.85 ± 0.01. GOLD, VINA, and AUTODOCK performed roughly similarly and acquired AUC values between 0.47 ± 0.03 and 0.88 ± 0.02. Here, DOCK performed fractionally worse than the other software—it failed with RXRα and MR as the AUC values ranged from abysmal to only barely above random (Table 2). However, it is also noteworthy that DOCK performed exceptionally well with NEU (AUC: 0.82−0.83 ± 0.02; Table 2; Figure 3). The docking success of GLIDE depended highly on a target and the used screening mode: With the HTVS mode, the AUC values ranged from unsuccessful (MR: 0.48 ± 0.03) to moderate (RXRα: 0.68 ± 0.03), and with the SP mode AUC, values remained comparable to other docking software. The hydrophobic, static, and tight binding pocket of MR and the large and polar binding pocket of PDE5 (Figure 2A, E) were especially problematic for docking with the test sets (Table 2).

Moreover, the high early enrichment values, EFd 1% and EFd 5%, produced by docking screenings do not necessarily translate into high AUC values and *vice versa*. To put it simply, a high AUC value relates to good overall enrichment, and the high early enrichment values indicate that more active compounds are found at the very beginning of the ranking list. Accordingly, in the case of NEU, PLANTS docking produced a relatively low

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

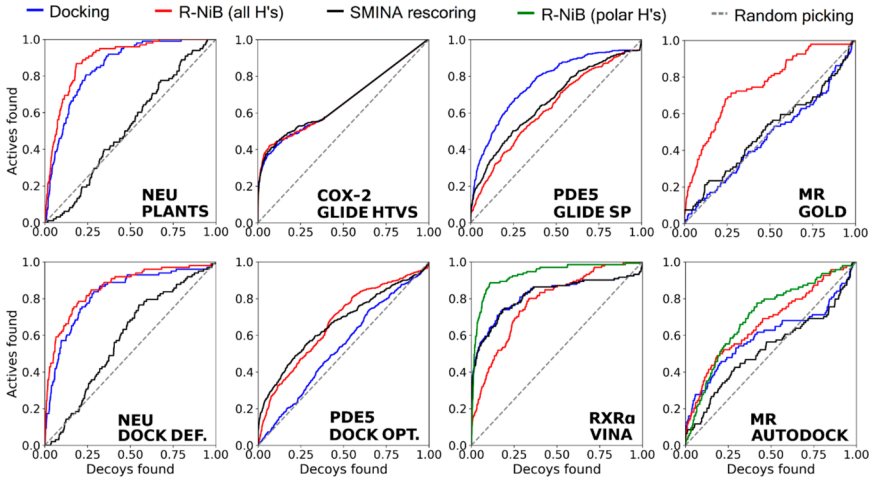


Figure 3. Docking and rescoring performance as receiver operator characteristic curves. The linear receiver operator characteristic (ROC) curves are plotted for the original docking and rescoring results of negative image-based rescoring (R-NiB; Figure 1)[21] or SMINA[45] rescoring. Benchmarking is shown for a selected assortment of results, but the full set of data is given in the Supporting Information for each software and test set in the form of linear (Figures S1−S5) and semilogarithmic ROC curves (Figures S6−S10). The figure was created using ROCKER0.1.4.[52]

EFd 1% value of 4.1, although the generated AUC value was a whopping 0.85 ± 0.01, indicating good overall enrichment (Table 2; Figure 3). In contrast, DOCK was able to produce an EFd 1% of 8.4, while AUC remained at 0.52 ± 0.03 for the RXRα test set. Generally, GLIDE and VINA were producing slightly better early enrichment than the other docking software (Table 2; Figure 3).

Notably, GLIDE and DOCK failed systematically to dock many of the active ligands and inactive decoys (Table 2). The software skipped most of the compounds in certain test sets regardless of their potential activity. In the case of GLIDE, the much faster HTVS screening mode discarded more ligands than the SP mode: the ability to dock active molecules of MR and NEU varied from 13% to 96% in the HTVS mode and from 34% to 100% in the SP mode, respectively. Likewise, DOCK was managing to dock far fewer compounds than the other algorithms (Table 2). With RXRα, only 26% of the active ligands were docked using the default settings. However, after tweaking the orientation and iteration number settings together with the conformer score cutoff, the number of docked ligands could increase slightly or even substantially, and 73% of actives could be docked, for example, with RXRα. Although only PLANTS, GOLD, and VINA were able to dock all active and decoy molecules, GLIDE and DOCK skipped a remarkably higher number of active molecules than any other tested software.

If omitting the fact that GLIDE docking skipped a lot of active compounds, the algorithm produced very impressive enrichment metrics. For example, AUC, EFd 1%, and EFd 5% values for MR were 0.88 ± 0.03, 50.0, and 58.3 with GLIDE HTVS, respectively, if the skipped molecules were ignored from the calculations (Table S1). In the case of RXRα, the corresponding values were 0.98 ± 0.01, 68.7, and 88.1, respectively. However, skipping over a half of the molecules can hardly be counted as a success, and when considering the skipped molecules, AUC, EFd 1%, and EFd 5% values decreased to 0.48 ± 0.03, 7.4, and 9.6 for MR, and 0.68 ± 0.03, 44.3, and 50.4 for RXRα, respectively (Table 2).

2.3. Negative Image-Based Rescoring Boosts Docking Performance. R-NiB (Figure 1) works well with different docking softwares and targets based on benchmarking (Table 3). Regardless, it is also clear that some docking algorithms benefit more from cavity-based rescoring than others (Table 3; Figure 3; Figures S1−S10). The rescoring works particularly well with DOCK, PLANTS, and GOLD. PLANTS docking results were improved here even more than in the prior study[21] by implementing slightly different PANTHER settings (Table S2). Simply put, these docking algorithms were consistently able to output several alternatives and high-quality docking poses for each test set compound, which is a prerequisite for improving the yield postdocking using R-NiB or other rescoring methods such as SMINA.[45] R-NiB works relatively well also with VINA and AUTODOCK, but the rescoring process itself is somewhat laborious due to the properties of these two software tools as described in Section 5.3.

VINA and AUTODOCK output the docking poses with only polar protons included. As the lack of nonpolar protons could affect negatively the R-NiB performance, the docked molecules were used in the rescoring either directly (polar H's in Table 3) or after adding nonpolar protons and MMFF94[53] partial charges (all H's in Table 3). The effects of this postprocessing to the R-NiB results varied (Table 3; Figure 3). For example, AUTODOCK and VINA performed exceptionally well with RXRα (Figures 1 and 2), and moreover, the cavity-based rescoring improved only the AUC and EFd 5% values without the added polar protons. However, this improvement was remarkable as the AUC and EFd 5% values increased notably with VINA, for example, from 0.82 ± 0.02 to 0.93 ± 0.02 and from 55.7 to 71.8, respectively. VINA docking for COX-2 was not improved by R-NiB, whereas the improvement was clear for AUTODOCK. With MR, VINA docking was improved more when all protons were added for the ligand conformers, whereas only the AUC values were improved when rescoring the AUTODOCK poses. Rescoring of the docking poses of NEU outputted by AUTODOCK and VINA was successful, but the improvement was higher when the ligands contained all protons.

From a practical standpoint, GLIDE emerges as the most problematic docking software of the bunch for R-NiB (Figure 1). Although the rescoring of the HTVS results improved enrichment for the targets such as RXRα and NEU, the rescoring of the SP poses improved the enrichment meaningfully only with NEU (Table 3). Depending on the test set, particularly, the

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Table 3. Performance of Negative Image-Based Rescoring[a]

			GLIDE			DOCK		VINA		AUTODOCK	
		PLANTS	HTVS	SP	GOLD	Def.	Opt.	Polar H's	All H's	Polar H's	All H's
RXRα	AUC	***0.82 ± 0.02***	0.68 ± 0.03	0.83 ± 0.02	***0.93 ± 0.02***	0.45 ± 0.02	***0.75 ± 0.02***	***0.93 ± 0.02***	0.78 ± 0.02	***0.94 ± 0.01***	0.80 ± 0.02
	EFd 1%	***21.4***	***47.3***	61.1	***62.6***	***19.1***	***48.9***	29.0	15.3	29.0	16.8
	EFd 5%	***40.5***	49.6	74.0	***82.4***	***20.6***	***61.8***	***71.8***	26.7	***77.1***	32.8
COX-2	AUC	***0.79 ± 0.01***	0.66 ± 0.01	0.73 ± 0.01	***0.76 ± 0.01***	***0.64 ± 0.01***	***0.72 ± 0.01***	0.74 ± 0.01	0.75 ± 0.01	***0.72 ± 0.01***	***0.73 ± 0.01***
	EFd 1%	***17.9***	24.1	31.0	***23.4***	***23.7***	***19.5***	11.3	11.0	***12.4***	***10.5***
	EFd 5%	***37.7***	***39.3***	***48.3***	***44.6***	***35.6***	***40.5***	34.5	32.0	***34.7***	***30.1***
PDE5	AUC	0.76 ± 0.01	0.62 ± 0.02	0.63 ± 0.02	0.70 ± 0.01	***0.65 ± 0.02***	***0.67 ± 0.02***	***0.79 ± 0.01***	***0.76 ± 0.01***	***0.73 ± 0.01***	***0.70 ± 0.01***
	EFd 1%	10.6	4.0	6.8	***10.3***	***4.3***	***3.8***	***13.3***	8.8	***10.8***	***8.5***
	EFd 5%	27.1	15.3	15.3	20.4	***20.6***	***18.3***	***27.6***	***23.1***	***20.6***	***17.1***
MR	AUC	***0.73 ± 0.03***	0.48 ± 0.03	0.50 ± 0.03	***0.76 ± 0.03***	0.47 ± 0.03	***0.55 ± 0.03***	***0.71 ± 0.03***	***0.68 ± 0.03***	***0.71 ± 0.03***	***0.67 ± 0.03***
	EFd 1%	***12.8***	7.4	6.4	***12.8***	0.0	0.0	2.1	***13.8***	3.2	8.5
	EFd 5%	***26.6***	***12.8***	18.1	***25.5***	***9.6***	***9.6***	***12.8***	***26.6***	13.8	22.3
NEU	AUC	***0.89 ± 0.02***	***0.82 ± 0.03***	***0.92 ± 0.02***	***0.93 ± 0.02***	***0.86 ± 0.02***	***0.88 ± 0.02***	***0.80 ± 0.03***	***0.81 ± 0.03***	***0.73 ± 0.03***	***0.77 ± 0.03***
	EFd 1%	***13.3***	***40.8***	***46.9***	***37.8***	***23.5***	***22.4***	***2.0***	***13.3***	***4.1***	***24.5***
	EFd 5%	***42.9***	***58.2***	***73.5***	***75.5***	***46.9***	***48.0***	***15.3***	***29.6***	***28.6***	***49.0***

[a]Bolded and in italics if better than the original docking results.

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

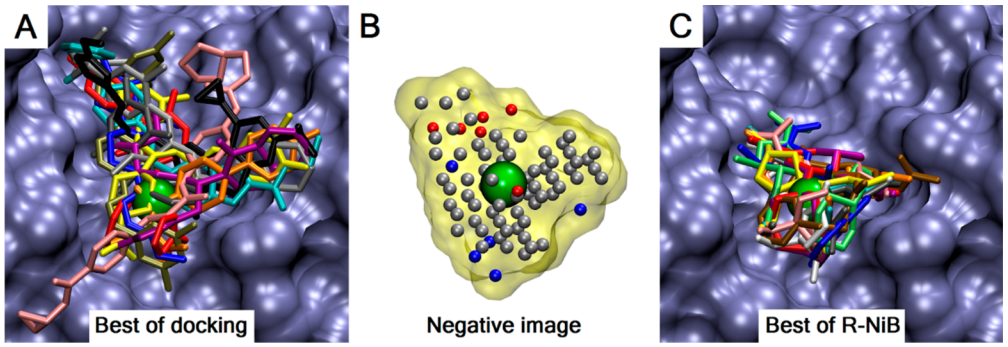


Figure 4. Negative image-based rescoring refocuses neuraminidase docking. (A) Best poses of 10 top-ranked docked compounds (stick models), sampled and selected by docking algorithm PLANTS,[37] are relatively scattered in the partially open surface pocket of neuraminidase (NEU; opaque magenta surface). The cavity center (green sphere), which is the geometric centroid of the co-crystallized ligand BANA206 (PDB: 1B9V),[50] was used to center PLANTS docking and NIB model generation with PANTHER.[22] (B) The NIB model (transparent yellow surface), which was used in the cavity-based rescoring with ShaEP,[32] is shown with the centroid. (C) Best poses of 10 top-ranked docked compounds selected by negative image-based rescoring (R-NiB; Figure 1)[21] form a much tighter cluster than scattered ligands/poses selected at the top by default docking scoring of PLANTS. The figure was created using BODIL[35] and Visual Molecular Dynamics or VMD 1.9.2.[36] See Figure 2 for more information.

GLIDE HTVS mode produced far fewer docking poses than the other docking software (Table S3). Accordingly, not only with the case of GLIDE but also other docking software often generated less conformers for the active ligands than for the decoy compounds. Although GLIDE skipped a varying number of molecules during docking and might generate a low number of conformers, the docked molecules were usually well separated to the active and inactive categories, and essentially, there was not much to perform the rescoring with (Table 2; Table S1).

2.4. Negative Images Refocus Docking. Both flexible docking and NIB model generation were performed using the same centroid coordinates; thus, the ability of the methods to focus their ranking on those compounds close to the center could be compared (Table S4). The more volume overlaps there are with the docked ligands and the template NIB model in the rescoring the higher the ShaEP scores, and as a result, the best-ranked poses are also more centered than the lower ranked ones. However, the best-ranked poses favored by the docking algorithms are also close to the cavity center, and overall, the data do not suggest that the rescoring would get its power from discarding ligands that are docked away from the cavity center. When the ligand-binding site is not a well-defined cavity (e.g., NEU in Figures 2B and 4), flexible docking typically generates spatially scattered poses for scoring (Figure 4A). In such a case, R-NiB focuses the compound selection effectively to the cavity volume of interest (Figure 4B, C), but the benefits of this refocusing are case specific (PDE5 vs NEU in Tables 2 and 3).

In addition to limiting the sampling to a certain cavity area or volume with the center and radius options, the docking scoring functions can also include steric terms that estimate the shape complementarity between the ligand poses and residues lining the cavity. Because R-NiB improves the docking yield mainly based on the shape similarity (see below) between the ligands and the cavity-based model, the capacity of default docking scoring to exploit this important part of molecular recognition was probed. The comparison of the ligands ranked best by the docking software against the cavity-based negative images indicates that the default scoring functions do not prefer ligands or their conformers that would best match the cavity shape as described by the PANTHER-generated NIB models. However, one must remember that the NIB models do not necessarily project the optimal dimensions of the cavity (Figure 2), and for this reason, these results regarding docking scoring are suggestive only (data not shown).

2.5. Shape Similarity Is Vital for Negative Image-Based Rescoring. ShaEP[32] compares both the shape and electrostatic potential (ESP) between the cavity-based NIB model and the docked molecule when calculating the total similarity score into the range from 0 to 1. By default, an equal 50/50 weight is given for the shape and ESP in the total scoring, and typically, it works well in ligand-based screening, NIB screening, and R-NiB[21,22,31] (Figure 1). As a rule, the shape similarity always gets a higher score than the ESP similarity, which makes the shape similarity between the docked ligand and the NIB model the determinant factor of R-NiB scoring.

The AUC and EFd values of the shape and ESP similarity scores were calculated separately for each test set and individual docking solution using R-NiB (Table S5). Here and in the other NIB studies,[21,22,31,54] the best shape similarity score reported for a molecule was typically two times higher in comparison to the best ESP similarity score. As a result, the impact of the shape similarity score on the total score (shape + ESP) has a major role with rescoring. Because the ligand-binding pocket of MR is highly hydrophobic (Figure 2A), R-NiB would improve the early enrichment of docking even more if the ESP similarity was forfeited altogether at least with rescoring poses outputted by GLIDE SP, GOLD, VINA, and AUTODOCK (Table S5). This also explains why the MR test set works so well with R-NiB but is more problematic for the standard docking functions that seem to underestimate the importance of shape complementarity. In the case of some docking software, the early enrichment can indeed be improved when completely ignoring the ESP; however, in none of these cases, AUC was improved. The ESP score worked particularly poorly in the case of GLIDE, VINA, and AUTODOCK.

Recently, promising results were reported for predicting activity by relying solely on the ESP similarity between the small molecules and the ligand-binding cavity, albeit the used data sets were limited in size, and importantly, no customary decoy predictions were processed.[55] Also, in R-NiB, the shape similarity scoring alone is usually not enough to acquire the best R-NiB enrichment, but a small push from ESP is needed for producing the top ranking. When considering only ESP scoring, the enrichment of the test sets remains too low for improving the docking yield. Nevertheless, occasionally ESP scoring can

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Table 4. Performance of SMINA Rescoring[a]

			GLIDE			DOCK			
		PLANTS	HTVS	SP	GOLD	Def.	Opt.	VINA	AUTODOCK
RXRα	AUC	***0.80 ± 0.02***	0.66 ± 0.03	0.77 ± 0.02	***0.78 ± 0.02***	***0.43 ± 0.02***	***0.67 ± 0.03***	0.82 ± 0.01	0.82 ± 0.02
	EFd 1%	***40.5***	32.8	41.2	***38.9***	***11.5***	***32.8***	38.9	35.9
	EFd 5%	***56.5***	41.2	55.7	***54.1***	***14.5***	***41.2***	54.2	54.2
COX-2	AUC	***0.76 ± 0.01***	0.66 ± 0.01	0.73 ± 0.01	***0.75 ± 0.01***	***0.64 ± 0.01***	***0.69 ± 0.02***	0.77 ± 0.01	***0.75 ± 0.01***
	EFd 1%	***34.0***	22.3	21.4	***32.6***	***20.0***	***19.5***	33.3	***24.6***
	EFd 5%	***48.3***	***37.5***	43.0	***47.6***	***29.9***	***34.5***	47.8	***42.3***
PDE5	AUC	0.70 ± 0.01	0.62 ± 0.02	0.66 ± 0.02	0.73 ± 0.01	***0.63 ± 0.02***	***0.66 ± 0.02***	0.66 ± 0.02	***0.71 ± 0.01***
	EFd 1%	***16.1***	***10.3***	***11.1***	***19.6***	***12.8***	***15.1***	***14.1***	***13.3***
	EFd 5%	25.6	20.6	21.4	30.4	***23.1***	***25.4***	***24.4***	***22.6***
MR	AUC	0.51 ± 0.03	0.48 ± 0.03	0.48 ± 0.03	0.51 ± 0.03	0.46 ± 0.03	***0.51 ± 0.03***	0.53 ± 0.03	0.53 ± 0.03
	EFd 1%	***7.4***	6.4	9.6	***7.4***	***4.3***	***4.3***	7.4	6.4
	EFd 5%	9.6	6.4	11.7	8.5	***8.5***	***11.7***	9.6	9.6
NEU	AUC	0.52 ± 0.03	0.46 ± 0.03	0.57 ± 0.03	0.60 ± 0.03	0.58 ± 0.03	0.60 ± 0.03	0.52 ± 0.03	0.47 ± 0.03
	EFd 1%	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
	EFd 5%	1.0	0.0	0.0	3.1	2.0	3.1	1.0	2.0

[a]Bolded and in italics if better than the original docking results. Underlined if also better than the negative image-based rescoring (R-NIB; Figure 1) protocol.

produce better results than a shape score alone; for example, this was the case with RXRα and GOLD docking (Table S5). This shape-centricity of R-NiB scoring (Table S5) is apparent for all targets and docking algorithms. Although a high level of shape similarity is a must for R-NiB implementation, a successful NIB model generally pertains also to some ESP similarity with the active ligands, reflecting the H-bonding capabilities of the target's cavity (Figure 1).

Benchmarking was also performed with ShaEP by aligning the co-crystallized ligands against *ab initio*-generated ligand conformers (Table S6). This standard ligand-based screening approach worked moderately well with some of the targets (e.g., COX-2: AUC 0.71 ± 0.01; EFd 1% 19.3; and EFd 5% 34.0), which highlights the limited diversity of the DUD-E test sets regarding shape similarity (Figure S11).[56−58] However, this does not mean that R-NiB is only able to find structurally very similar molecules: the methodology does not fare any worse in finding different molecule clusters in comparison to original PLANTS scoring or SMINA[45] rescoring (Section 2.7).

2.6. Rescoring Mixed Ligand Sets Using Negative Images. PDE5 is a demanding nut to crack for the R-NiB methodology (Figure 1)[21] or for the flexible docking algorithms (Table 2). This difficulty arises from the fact that the known PDE5 inhibitors are a very diverse set of ligands, such as exhibited by sildenafil or tadalafil[47,48] (Figure S12), whose binding poses and, ultimately, binding locations inside the cavity vary considerably. Furthermore, the PDE5 binding pocket is large (Figure 2E), and the ligand binding is affected by water coordination and induced-fit effects. Indeed, clustering based on Daylight's fingerprint and Tanimoto similarity indicates that the PDE5 test set contains chemically more diverse active ligands than the other tested sets (Figure S11).

Due to these effects, R-NiB was performed using two different NIB models for PDE5 that were created based on the PDB entries co-crystallized with either sildenafil or tadalafil.[47,48] Rescoring using neither one of the models alone produced significant improvement in comparison to the original docking (Table S7). When the two NIB models were used together, the results improved only moderately in comparison to the single NIB model rescoring (Table 3). Although PDE5 early enrichment was slightly improved for GOLD docking and the very early enrichment improved with PLANTS as well, the AUC values generally decreased as a result of the two-model R-NiB treatment for PLANTS, GOLD, and GLIDE. Although two models might provide a more comprehensive picture of the flexible PDE5 cavity space than a single model, the individual weight of the templates is challenging to assign for R-NiB scoring. The scale of the ShaEP score is directly affected by the amount of the atoms present in the NIB models or ligands being compared.

R-NiB improved the PDE5 enrichment, especially for DOCK, VINA, and AUTODOCK that managed worse with the demanding test set. With DOCK, failing to separate the active molecules from the inactive decoys for the PDE5 set by default (Table 2; Figure 3), the cavity-based rescoring provided easily a minimal improvement (Table 3). In fact, the excellent R-NiB enrichment suggests that the sampling of DOCK works far better than its default scoring function. The rescoring of AUTODOCK and VINA docking poses using R-NiB improved both AUC and early enrichment values of PDE5, especially when the docked molecules contained only the polar protons (Table 3). However, with the fully protonated docking poses (nonpolar protons included), early enrichment could not be improved for VINA using R-NiB. Rescoring of PDE5 docking results was more fruitful with SMINA than with R-NiB (see below).

2.7. Negative Image-Based Rescoring vs SMINA Rescoring. R-NiB (Figure 1) performance has been compared against the rescoring algorithm XSCORE[59] previously.[21] Because R-NiB outperformed XSCORE at least with the default settings, cavity-based rescoring methodology is now set against another rescoring algorithm called SMINA.[45] Here, this software was used only for docking rescoring using its default empirical scoring function. On the face of it, SMINA seems like a

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Table 5. Root-Mean-Square Deviation: Docking and Rescoring vs X-ray Crystallography

			RMSd < 1.0 Å		RMSd < 2.0 Å		RMSd < 3.0 Å	
Docking software	Scoring method	Docked/co-crystallized[a]	Best scored[b]	All[c]	Best scored[b]	All[c]	Best scored[b]	All[c]
PLANTS	default	31/31	8	15	15	27	**18**	29
	R-NiB		**9**		**16**		**18**	
	SMINA		5		8		12	
GLIDE HTVS	default	26/31	9	11	12	17	17	20
	R-NiB		**11**		**15**		**19**	
	SMINA		8		12		16	
GLIDE SP	default	29/31	**12**	16	17	21	21	24
	R-NiB		11		**18**		**23**	
	SMINA		11		16		**23**	
GOLD	default	31/31	12	15	17	23	19	24
	R-NiB		11		**21**		**23**	
	SMINA		**13**		17		21	
DOCK def.	default	24/31	7	13	11	14	12	15
	R-NiB		7		10		10	
	SMINA		7		**12**		**13**	
DOCK opt.	default	31/31	**8**	12	**14**	19	**15**	22
	R-NiB		**8**		**14**		14	
	SMINA		6		13		**15**	
VINA	default	31/31	**9**	12	**19**	26	**21**	28
	R-NiB		**9**		17		18	
	SMINA		7		9		12	
AUTODOCK	default	31/31	**10**	13	**17**	21	19	24
	R-NiB		9		**17**		**21**	
	SMINA		8		13		16	

[a]Number of docked active ligands with known poses from the X-ray crystallographic studies. [b]Root-Mean-Square deviation (RMSd) value of the best pose suggested by the docking software, negative image-based rescoring (R-NiB), or SMINA. [c]Best RMSd value, if all docking poses are compared against the co-crystallized ligand conformer. The scoring results with most matches with the verified poses are bolded. The representative PDB codes for the used X-ray crystal structures were the following: 4K6I, 1FM9, 1RDT, 1MVC, 4K4J, and 3A9E for the retinoid X receptor alpha (RXRα); 4PH9, 4M11, 3QMO, 5KIR, 1PXX, 3NT1, 5JVZ, and 4COX for cyclooxygenase-2 (COX-2); 3TGE, 3HC8, 1XOZ, and 1TBF for phosphodiesterase-5 (PDE5); 5MWY, 3VHU, 2AA5, 2AA2, and 4UDA for the mineralocorticoid receptor (MR), and 1B9V, 1XOG, 2QWE, 1A4Q, 2QWG, 1LTF, and 6HCX for neuraminidase (NEU).

worthy competitor to R-NiB because it is not only fast but also relatively easy to use. However, benchmarking indicates that the success of SMINA is more case-specific than that of R-NiB (Table 4; Figure 3).

SMINA did not improve docking enrichment for NEU at any level. MR was also difficult for SMINA rescoring except for solutions output by DOCK (Table 4). SMINA performed particularly well with PDE5 as it was able to improve at least the early enrichment of 1% even better than R-NiB with every docking software (Table 4; Figure 3). If not including MR and NEU test sets, SMINA worked well with DOCK, PLANTS, and GOLD. On the other hand, it was far less successful with GLIDE and VINA because it was able to improve the early enrichment only occasionally. As a fork of VINA, it was unexpected that SMINA rescoring of VINA docking results proved to be problematic. With AUTODOCK, SMINA was able to improve not only PDE5 but also COX-2 results and with a higher yield than R-NiB. In addition, SMINA outperformed R-NiB in some other cases, such as RXRα and COX-2 sets docked with PLANTS and the COX-2 set docked with GOLD.

All in all, when SMINA rescoring seemed to work out, the yield improvement was notable. SMINA was also consistently able to increase the very early enrichment rather than the AUC values (Figure 3; Figures S1−S10; Table 4).

2.8. Docking Predictions vs X-ray Crystallography. In addition to benchmarking (Table 2; Figures S1−S10), docking or R-NiB (Figure 1) performance can be evaluated by comparing the predicted and/or top-ranked poses against the co-crystallized protein-bound ligand conformers. These comparisons can be simple on-screen 3D visualizations, but numerically, they are typically processed as the Root-Mean-Square deviation (RMSd) values. These kinds of comparisons have been done widely with different docking software.[7,8,10,11] Recently, PLANTS was stated to be the best software when considering the ability of a docking algorithm to reproduce the co-crystallized ligand binding pose and rank it as the best pose.[9] However, because the results vary depending on the survey and the used test sets, the ultimate ranking between docking algorithms remains elusive.

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Generally, docking is regarded as successful, if the RMSd value is below either 1.5 or 2.0 Å; however, a larger threshold is sometimes needed to detect truly worse or better docking poses. Namely, the smaller the ligand is, the higher the impact of a minor (and potentially trivial) difference between the predicted and verified poses is to cause a relatively large RMSd shift. Moreover, the ligand could have more than one valid binding mode, and only one of them is co-crystallized with the protein in the X-ray crystal structure.[12] Even so, if one has a high-resolution ligand−receptor complex 3D structure available, the comparison against the experimental structures should be performed by default. In total, 31 active molecules in the five test sets were found to have representative X-ray crystal structures available in PDB (6 × RXRα, 8 × COX-2, 4 × PDE5, 5 × MR, and 8 × NEU; Table 5).

When focusing solely on the best poses selected by the docking software themselves, VINA found 61% highly similar poses in comparison to the co-crystallized ligand conformers (RMSd < 2.0 Å in Table 5). If considering only the poses that have RMSd values of <1.0 Å with the co-crystallized conformers, GLIDE SP and GOLD outperformed the other software by predicting 39% of the binding poses correctly. The MR binding poses were the easiest ones to reproduce by the tested docking software (Table 5), and not surprisingly, the reproduction of the PDE5 inhibitor binding poses turned out to be the most demanding case for the docking algorithms.

For most software, more than one alternative docking pose could be outputted. From the rescoring point of view, this feature is relevant because any of these outputted poses (not just the highest ranked ones) could be the proper bioactive conformation of the molecule that is sought after. Thus, the RMSd values were also calculated for the alternative conformers (column "All" in Table 5) and not just the best-ranked poses. PLANTS and VINA outperformed the other software by reproducing 87% and 84% of the co-crystallized poses, respectively (RMSd < 2.0 Å in Table 5). When focusing only on those poses with the RMSd values of <1.0 Å, GLIDE SP was slightly better than PLANTS or GOLD because it reproduced 52% of the co-crystallized poses as opposed to 48%.

Because the cavity-based rescoring improves the docking results, sometimes substantially (Table 3), it seems logical that R-NiB (Figure 1) would also focus on those ligand conformers that match the verified binding modes. However, the comparison against the co-crystallized ligand conformers indicates that R-NiB was only slightly better in selecting the correct poses in comparison to the original docking algorithms (Table 5). The best-rescoring matches were acquired with GOLD and GLIDE HTVS as R-NiB selected three and four more molecules, respectively, with the RMSd threshold of <2.0 Å compared to the original docking scoring. Notably, PLANTS reproduced altogether 27 of the 31 verified ligand conformers with the RMSd value of <2.0 Å. Despite this good premise, R-NiB could pick correctly only 16 of these high-quality poses outputted by PLANTS. SMINA performed worse than R-NiB, as it typically failed to select more of the co-crystallized poses with any of the tested thresholds than the original docking (Table 5).

3. DISCUSSION

Although the importance of the shape complementarity for molecular recognition has been acknowledged for a long time,[60,61] the meaningful use of this revelation in drug discovery has proven difficult. It is easy to fathom the importance of shape similarity between the receptor's cavity and a high-affinity ligand, but the exact dimensions of the relevant pocket are a lot harder to ascertain or utilize in practice. The importance of shape constraints in docking has been noted in recent D3R Grand Challenges,[62,63] and importantly, several novel methods address the issue by applying experimental ligand shape constraints,[64,65] template matching,[66] or interaction fingerprint matching[67] to improve docking accuracy. The negative image-based rescoring (R-NiB; Figure 1) provides a tangible way to address this issue in the docking-based virtual screening assays.[21,22,31]

Negative images are an intuitive way to abstract and visualize the shape/electrostatics features of the ligand-binding pockets/cavities, grooves, or even tunnels present in the protein 3D structures (Figure 2). Therefore, various cavity detection or analysis algorithms such as VOIDOO/FLOOD,[68] FPOCKET,[69] CAVER,[70] POVME,[71−73] and SITEMAP[74,75] have been developed to perform cavity visualization or druggability, accessibility, size/volume, or flexibility estimation. PANTHER[22] differs from the other software as it is not primarily focused on the analysis, visualization, or even binding site detection or prediction, but instead, it was intended to facilitate cavity-based rigid docking or negative image-based (NIB) screening.[22,25,30]

In other words, PANTHER-generated NIB models not only mirror the shape/electrostatics of the binding pocket but also their filler atom/cavity point and charge composition were intended to be drug-like from the start (Figure 2). The volume of the best NIB model does not necessarily cover the entire ligand-binding cavity, but it preferably spans areas that facilitate drug binding (Figure 2; Figure S12). Precisely, due to this drug-likeness, the NIB models can be used directly as pseudoligand templates in the similarity comparison with explicit docking poses in R-NiB (Figure 1). The model generation has been shown to work successfully without prior structural data on the ligand binding;[21,54] however, the co-crystallized ligands can be used to improve the model dimensions especially regarding early enrichment (Figure 2; Figure S12). The user adjusts the model via the PANTHER settings (Table S2) to ensure its drug-like composition, and by doing so, takes it further away from simply depicting the cavity.

The premise of the R-NiB methodology (Figures 1 and 2) is two-fold. First, flexible docking is expected to sample correctly the bioactive ligand binding poses, even if the default scoring falters in the ranking of those same poses.[14,17,18] Second, ligand−receptor complex formation relies heavily on shape complementarity,[60,61] and consequently, the best docking poses could be recognized by putting this facet of molecular recognition into focus. Indeed, the results indicate that R-NiB boosts docking by simply comparing the alternative docking poses against the shape/charge of the binding cavity's negative image (Figures 1 and 2). The performance boost comes mainly from the shape similarity between the ligand-binding cavity and the docked ligands (Figure 2; Table S5). R-NiB works not only with all six tested docking softwares (Table 3) but also with very different drug targets (Figure 2; Table 1). When also considering the rapid calculation times,[21] R-NiB is clearly a handy tool for improving the efficiency of docking-based high-throughput virtual screening (HTVS) assays.

Docking sampling is typically performed within a loosely defined volume given either as a 3D box or sphere. The protein-bound ligands can assist in defining the search area; however, flexible docking brings out binding poses that are a lot more

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

scattered than what the negative images in R-NiB allow (Figure 4). As a result, the docking algorithms go on to score and rank all docking poses of which some might be noticeably off the cavity center or binding hot spot(s), in an equal manner. This lack of focus can be problematic, for example, when docking ligands into cavities located on the protein surfaces such as is the case with neuraminidase (NEU; Figures 2B and 4). In this context, R-NiB provides the needed focus by limiting the compound selection to specific locations/volumes of the protein's cavity (Figure 4). If the model generation is confined to a certain subsection of the cavity, the rescoring also becomes specialized or focused to a certain subset of the screened active compounds.[54]

If the target's pocket is spacious, it is advisible to generate alternative NIB models to study the different cavity parts separately instead of trying to build one model that fits all the parts. Moreover, a single static protein 3D conformation is frequently not enough to depict the ligand-binding cavity, when benchmarking diverse ligand sets or attempting to discover truly novel hit compounds using flexible docking.[13,76,77] This is because the reciprocal induced-fit effects between the ligand and its receptor can lead to profoundly different binding modes that cannot be predicted using a single protein conformation. For example, a single NIB model cannot distinguish adequately two or more distinct ligand types included in phosphodiesterase-5 (PDE5; Figure S12) or estrogen receptor alpha test sets.[21] Instead of trying to build an all-inclusive NIB model using a single input structure, docking and rescoring should be performed using several target protein conformations.

An interesting *in silico* solution for this multiconformation problem is to sample the protein's conformational ensemble using molecular dynamics (MD) simulations prior to the docking/rescoring or optimization of the postdocking complex states using energy minimization.[78] Although the MD simulations can be very useful in refining the docking-based binding mode predictions,[79−85] the abundance of stochastic noise, i.e., too many random and irrelevant changes, complicates or even prevents their effective HTVS usage. On the other hand, the short minimizations of the complexes cannot necessarily overcome completely wrong configurations or distinguish docking- or MD-related artifacts. Moreover, the postprocessing works reliably only after performing the initial docking and cavity-based rescoring because both docking and R-NiB are sensitive regarding input protein conformation.

SMINA[45] is an interesting rescoring software due to its speed, ease of use, and possibility of generating custom scoring functions. However, at least using its default settings, SMINA was more case-specific than R-NiB. SMINA outperformed R-NiB with PDE5 but also in some cases with COX-2. The COX-2 test set contains both selective and nonselective ligands,[39] which likely affects the R-NiB results when using only a single NIB model similar to the estrogen receptor alpha test set.[21] Overall, R-NiB worked better than SMINA in rescoring at least with these specific DUD-E test sets (Table 3 vs Table 4) with the notable exception of very early enrichments (Figures S6−S10). By building specific scoring functions, SMINA could work better with specific targets; however, further tinkering of the NIB models would improve the R-NiB results as well. In a prior study,[21] R-NiB was performed using slightly different default settings of PANTHER, and implementation of only minor changes led to improvement of the PLANTS rescoring results (Table S2).

The rescoring success is not dependent on the amount of outputted alternative docking poses but the quality of the generated ligand conformers. Although the quantity is not proportional to the quality, there need to be a few conformers, or at least one docking pose for each compound to perform any rescoring. Particularly, if the average number of docked conformers highly differs between active and decoy molecules, it distorts the original active/decoy ratio, and with high conformer numbers, the R-NiB success sometimes decreases (data not shown). Thus, a high number of outputted conformers does not guarantee success, and *vice versa*, R-NiB can improve the docking yield also if only a couple conformers are given (e.g., RXRα with GLIDE HTVS; Table 3).

R-NiB was not markedly better than default docking scoring in recognizing the experimentally verified ligand binding poses from the pool of alternative docking poses (Table 4). While bigger validation sets would have been preferable, these results already show that docking is generally able to sample the ligand binding correctly, if several conformers are outputted.[18] R-NiB also outperformed SMINA in recognizing the correct poses albeit it is not infallible. Despite these encouraging results, R-NiB is, above all, meant for HTVS usage instead of predicting individual ligand-binding poses. As docking has already positioned the molecules inside the binding pocket, R-NiB only selects those molecules with the key shape/electrostatics features. This means that the match between the NIB model and the docked active ligands cannot be optimal for any of them when boosting the docking screening yield—it is always a give-and-take situation.

4. CONCLUSIONS

Negative image-based rescoring (R-NiB; Figure 1) improved the performance of six popular docking softwares, including GLIDE, PLANTS, GOLD, DOCK, AUTODOCK, and VINA[34] (Table 1), in benchmark testing (Table 3). R-NiB improved the docking enrichment most consistently with PLANTS, GOLD, and DOCK. The direct shape/electrostatics comparison between the explicit docking poses and the cavity-based negative images (Figure 2) was made for five target proteins. Although there was software- and target-specific differences, both the overall enrichment and the early enrichment were improved by R-NiB in most cases (Table 2 vs Table 3). The results indicate that R-NiB is an excellent solution for rescoring flexible docking poses that are generated using a single high-quality protein 3D structure (Table 3; Figures S1−S10). In short, the cavity-based rescoring works well in benchmarking regardless of the used docking software, thus making R-NiB an attractive addition to any docking-based virtual screening assay.

5. EXPERIMENTAL SECTION

5.1. Ligand Preparation. Retinoid X receptor alpha (RXRα; Figure 1), cyclooxygenase-2 (COX-2), phosphodiesterase type 5 (PDE5), mineralocorticoid receptor (MR), and neuraminidase (NEU) small-molecule test sets (Table 1) were acquired from the DUD-E (A Database of Useful (Docking) Decoys − Enhanced) database.[39] Although very useful, some of the DUD-E sets (RXRα being a notable exception) have been criticized as containing, for example, shape similarity biases that can favor ligand-based screening.[56−58] The preparation of the ligands to the 3D SYBYL MOL2 format including the addition of tautomeric states, OPLS3[86] partial charges, and protonation at pH 7.4 was done with MAESTRO 2017-1 (Schrödinger, LLC,

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

NY, USA, 2017) as described in a prior study.[21] A single conformer was generated for each compound from a SMILES string; accordingly, the same ligand 3D conformer was used as the input for all docking algorithms. For AUTODOCK4.2.6[44] and AUTODOCK VINA1.1.2,[34] the ligands were converted to the PDBQT format using the prepare_ligand4.py script of AUTODOCKTOOLS1.5.6.[44]

5.2. Protein Preparation. The protein 3D structures[33,46−50] used in the docking and rescoring steps were solved using X-ray crystallography and acquired from the Protein Data Bank (PDB; Figure 2; Table 1).[87,88] For PLANTS[37] and GLIDE[41,42] docking, the PDB entry editing, including the PDB-to-MOL2 conversion and removal of the unnecessary protein-bound ligands was performed in the BODIL Molecular Modeling Environment.[35] REDUCE3.24[89] was used to protonate the structures at pH 7.4. For GLIDE docking, the default settings of the Protein Preparation Wizard in MAESTRO 2018-1 were used in the protein preprocessing, but the pH was set to match 7.4. For AUTODOCK and VINA docking, the PDB entries were converted to the PDBQT format and prepared using AUTODOCKTOOLS. The preparation included incorporation of the Gasteiger(-Marsili) partial charges,[90] addition of polar protons, and removal of crystal waters or extra co-crystallized ligands. The protein preparation for UCSF DOCK6.8[43] docking was performed with UCSF CHIMERA1.12[91] using the default settings of the Dock Prep tool. The protonation of histidines was set unspecified, and the protein surface was created using the Write DMS tool in CHIMERA.

5.3. Molecular Docking. The centroid coordinates of the protein-bound or co-crystallized ligands in the PDB entries (Figure 2) were used as the centers for the flexible docking simulations. If not specified otherwise below, the docking was performed using the default settings. At maximum, 10 alternative poses were set to be outputted for the rescoring, and the radii were set to 10.0 Å. In the case of AUTODOCK and VINA, the conversion of molecule files to the PDBQT format needs to be done separately, whereas the MOL2 and PDB formats are suitable for other programs. The example input files containing settings for docking are included in the Supporting Information.

1. GOLD5.6.3[40] uses a genetic algorithm optimizer in docking sampling and a force field-based scoring function GoldScore. The scoring function considers H-bonding energy, van der Waals energy, potential metal interaction, and torsional strain in the binding pose estimation.
2. GLIDE2018-1[41,42] uses a systematic search technique in the ligand placement and Emodel scoring function that combines a force field-based method with empirical scoring function GlideScore. The compounds were docked using the default settings with both high-throughput virtual screening (HTVS) and standard precision (SP). Extra precision (XP) was not used as it is not suitable for processing high numbers of molecules.
3. PLANTS1.2,[37] whose docking sampling relies on an ant colony optimization method, was used to dock the compounds included in the DUD-E sets using the default settings already in a prior study.[21] The scoring was performed using ChemPLP that combines the piecewise linear potential (PLP) and GOLD's ChemScore. PLP is used to model steric complementarity between the ligand and its protein receptor, whereas ChemScore is used to calculate hydrogen and metal bonding, ligand heavy-atom clashes, and internal energies.
4. AUTODOCK4.2.6[44] relies on a Lamarckian genetic algorithm for docking sampling and semiempirical free energy force field for scoring.[92] Grid box dimensions were adjusted to roughly match the 10 Å docking radius with a default 0.375 Å spacing. To be better suited for virtual screening, the maximum number of energy evaluations was decreased to 2,500,000 in the docking procedure.
5. AUTODOCK VINA1.1.2[34] is based on AUTODOCK, but the algorithm uses the Broyden−Fletcher−Goldfarb−Shanno (BFGS) quasi-Newton method for the conformation generation and a combination of empirical and knowledge-based scoring functions. The maximum number of binding modes was set to 10, and the search space was set to 17−20 Å (corresponds roughly to 10 Å docking radius) depending on the size of the docked ligand. As the output of VINA[34] docking score contains only one decimal by default, the poses were reprocessed so that the search space was omitted (*−score_only* option) to separate them better for ranking purposes.
6. DOCK6.8[43] uses shape-based ligand placement and a grid-based energy scoring, although other scoring functions and their combinations can also be employed. The space of the ligand-binding pocket is determined with the DOCK accessory SPHGEN. A minimum sphere radius of 1.0 Å was used. The spheres within 8−10 Å of the co-crystallized ligand were selected. The grid files were generated using a 17−20 Å grid box with a default of 0.3 Å grid spacing depending on the size of the binding pocket. The flexible docking, which outputted 10 conformers for each compound, was performed with the default settings, if successful. To decrease the number of skipped molecules, the maximum number of orientations, anchor orientations, and iterations per cycle per anchor was increased to 1000, and the conformer score cutoff was set to 200 kcal/mol.

5.4. Root-Mean-Square Deviation Calculations. The X-ray crystal structures of the target proteins with bound active ligands, if available, were acquired from PDB.[87,88] The co-crystallized ligands also included in DUD-E were located based on their specific SMILES (Simplified Molecular-Input Line-Entry System) strings and ChEMBL database[93] codes. If several 3D structures were available, the one with the best resolution was selected. The backbone C^{α} atoms of the protein structures containing the active ligands were superimposed with the template structure used in docking with VERTAA in BODIL.[35] The realigned coordinates of the co-crystallized ligands and the docked ligands were both converted to the MAE format using MAESTRO2018-1. The Root-Mean-Square deviation (RMSd) calculations were performed for the ligands with the rmsd.py script in MAESTRO. In the case of VINA and AUTODOCK, the best rescoring results were used for the RMSd evaluation.

5.5. Negative Image Generation. The negative images or NIB (negative image-based) models of the target proteins' ligand-binding cavities (Figure 2) were generated using PANTHER0.18.15.[22] The cavity centroids were calculated directly using the ligands present in inputing the protein structures (Figures 1-2; Table 1). The same default PANTHER settings were utilized for all models, if not specified otherwise. The NIB model volumes were limited using the ligand distance limit of 1.5 Å (2.0 Å for the NEU); i.e., the models were not

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

allowed to grow too far from the area occupied by the co-crystallized ligands. The box radius option was set to 20.0−27.0 Å depending on the target protein's cavity volume.

The face-centered cubic (FCC) packing method was used with RXRα, NEU, and tadalafil-bound PDE5 structures, whereas the less dense body-centered cubic lattice (BCC) packing was selected for COX-2, MR, and sildenafil-bound PDE5 structures. With COX-2, the lining id angle option was decreased to 10° to include one positively charged cavity point near the side-chain oxygen of Gln178 in the model. With MR, the radius for the charged atoms option was decreased to 0.2 Å, and the exclusion distance for the charged atoms and their residues was set to 0.4 Å. This was done to make one end of the NIB model slightly thicker. The only positively charged cavity point near the main chain oxygen of Asn306 in the NIB model of RXRα was removed from the negative image manually as it was considered unconnected with the rest of the model.

The PANTHER input files, PDB files, and NIB models are included in the Supporting Information.

5.6. Negative Image-Based Rescoring. The negative image-based rescoring (R-NiB; Figure 1)[21] or the similarity comparison of the shape/electrostatics between the docked poses and the cavity-based negative image was performed with ShaEP1.1.3.[32] The similarity comparison was done without superimposing the original docking poses with the template NIB model using the *−noOptimization* option. For rescoring, the docked molecules need to be protonated and contain partial charges. However, both AUTODOCK and VINA output docking poses lacking nonpolar protons with the Gasteiger-(−Marsili) partial charges[90] in contrast to the input molecules containing all protons and the user-selected charges. In these specific cases, the outputted conformers were protonated at pH 7.4, and the Merck Molecular Force Field 94 (MMFF94)[53] partial charges were incorporated with OBABEL2.4.1[94] before the rescoring.

5.7. SMINA Rescoring. SMINA[45] (November 9, 2017; based on AUTODOCK VINA1.1.2[34]) is a freely downloadable fork of VINA.[34] Although it maintains most of the VINA properties, SMINA was designed with energy minimization and scoring in mind. It has a default scoring function that emphasizes the steric term or shape similarity between the ligand and its receptor, but the user also has a possibility to create custom scoring functions with different scoring term weights.[1] To compare the performance of R-NiB (Figure 1)[21] against the default scoring function of SMINA, all the docking solutions outputted by the tested algorithms were also rescored using SMINA.

5.8. Figure Preparation and Data Analysis. Figures 1, 2, and 4 were prepared using BODIL[35] and VMD1.9.2.[36] The area under the curve (AUC) and the early enrichment factors (EFd) were calculated with ROCKER0.1.4,[52] which uses the Wilcoxon statistic[95] to estimate the standard deviation. The EFd values reported in this study correspond to the percentage of true positive ligands discovered when 1% or 5% of the decoy compounds have been found. The receiver operating characteristics (ROC) curves in Figure 3 were plotted with ROCKER.[52] If the docking software could not produce a docking pose for a compound, these molecules were added to the bottom of the results. In practice, both the skipped or undocked active ligands and decoy compounds were added in even ratios; however, the even distribution always started with a decoy. This practice makes the EFd and AUC values or ROC curves comparable between different software. Due to the heterogeneity of the experimental measurements for the DUD-E test sets, no activity correlation was estimated. To-be-released in-house algorithm SDFCONF0.8.20 was used to calculate the average geometric centroids for the top-ranked docking poses.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00383.

Figures S1−S5: Linear receiver operator characteristic curves. Figures S6−S10: Semilogarithmic receiver operator characteristic curves. Figure S11: Clustering of active ligands based on Daylight's Fingerprint and Tanimoto similarity. Figure S12: Alternative X-ray crystal structures of phosphodiesterase-5 in complex with two inhibitors. Table S1: Performance of docking algorithms in benchmarking when ignoring skipped compounds. Table S2: Effect of PANTHER settings for negative image-based rescoring of PLANTS docking results. Table S3: Average ligand conformer numbers from flexible docking simulations. Table S4: Distance of the 10% of top-ranked docking poses from the cavity center. Table S5: Effect of shape and electrostatics similarity on the negative image-based rescoring. Table S6: Performance of ligand-based screening. Table S7: Negative image-based rescoring of the phosphodiesterase-5 test set with two alternative cavity-based models. (PDF)

The zipped file contains PANTHER input files, PDB files of target proteins, NIB models, and docking input files for each software. (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pekka.a.postila@jyu.fi.

ORCID

Olli T. Pentikäinen: 0000-0001-7188-4016

Pekka A. Postila: 0000-0002-2947-7991

Author Contributions

S.T.K. performed the virtual screening experiments. S.L. assisted S.T.K. in the analysis. S.T.K. wrote the manuscript with the help from O.T.P. and P.A.P. S.T.K., O.T.P., and P.A.P. designed the experiments. P.A.P. supervised the study.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Turku University Foundation and Finnish Cultural Foundation are acknowledged for personal grant to S.T.K. The Finnish IT Center for Science (CSC) is acknowledged for generous computational resources (OTP; Project Nos. jyy2516 and jyy2585). Visual Molecular Dynamics (VMD) was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana−Champaign.

■ REFERENCES

(1) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 146−157.

(2) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9*, 91−102.

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

(3) Ekins, S.; Mestres, J.; Testa, B. In Silico Pharmacology for Drug Discovery: Methods for Virtual Ligand Screening and Profiling. *Br. J. Pharmacol.* **2007**, *152*, 9−20.
(4) Zoete, V.; Grosdidier, A.; Michielin, O. Docking, Virtual High Throughput Screening and in Silico Fragment-based Drug Design. *J. Cell. Mol. Med.* **2009**, *13*, 238−248.
(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(6) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking: Current Status and Future Challenges. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 15−26.
(7) Liu, Z.; Wang, R.; Li, X.; Cheng, T.; Li, Y. Evaluation of the Performance of Four Molecular Docking Programs on a Diverse Set of Protein-Ligand Complexes. *J. Comput. Chem.* **2010**, *31*, 2109−2125.
(8) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964−12975.
(9) Ren, X.; Shi, Y.-S.; Zhang, Y.; Liu, B.; Zhang, L.-H.; Peng, Y.-B.; Zeng, R. Novel Consensus Docking Strategy to Improve Ligand Pose Prediction. *J. Chem. Inf. Model.* **2018**, *58*, 1662−1668.
(10) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(11) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(12) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." Structure. *Structure* **2009**, *17*, 489−498.
(13) Jain, A. N. Effects of Protein Conformation in Docking: Improved Pose Prediction through Protein Pocket Adaptation. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 355−374.
(14) 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. D.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(15) Kolb, P.; Irwin, J. Docking Screens: Right for the Right Reasons? Curr. *Curr. Top. Med. Chem.* **2009**, *9*, 755−770.
(16) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear Hormone Receptor Targeted Virtual Screening. *J. Med. Chem.* **2003**, *46*, 3045−3059.
(17) Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742−755.
(18) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(19) Feinstein, W. P.; Brylinski, M. Calculating an Optimal Box Size for Ligand Docking and Virtual Screening against Experimental and Predicted Binding Pockets. *J. Cheminf.* **2015**, *7*, 1.
(20) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. Comparative Study of Several Algorithms for Flexible Ligand Docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.
(21) Kurkinen, S. T.; Niinivehmas, S.; Ahinko, M.; Lätti, S.; Pentikäinen, O. T.; Postila, P. A. Improving Docking Performance Using Negative Image-Based Rescoring. *Front. Pharmacol.* **2018**, *9*, 260.
(22) Niinivehmas, S. P.; Salokas, K.; Lätti, S.; Raunio, H.; Pentikäinen, O. T. Ultrafast Protein Structure-Based Virtual Screening with Panther. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 989−1006. Software can be found at www.medchem.fi/panther (accessed 12.12.2018).
(23) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(24) Virtanen, S. I.; Niinivehmas, S. P.; Pentikäinen, O. T. Case-Specific Performance of MM-PBSA, MM-GBSA, and SIE in Virtual Screening. *J. Mol. Graphics Modell.* **2015**, *62*, 303−318.
(25) Niinivehmas, S. P.; Virtanen, S. I.; Lehtonen, J. V.; Postila, P. A.; Pentikäinen, O. T. Comparison of Virtual High-Throughput Screening Methods for the Identification of Phosphodiesterase-5 Inhibitors. *J. Chem. Inf. Model.* **2011**, *51*, 1353−1363.
(26) Ahinko, M.; Niinivehmas, S.; Jokinen, E.; Pentikäinen, O. T. Suitability of MMGBSA for the Selection of Correct Ligand Binding Modes from Docking Results. *Chem. Biol. Drug Des.* **2019**, *93*, 522−538.
(27) 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.
(28) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(29) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2006**, *46*, 380−391.
(30) Virtanen, S. I.; Pentikäinen, O. T. Efficient Virtual Screening Using Multiple Protein Conformations Described as Negative Images of the Ligand-Binding Site. *J. Chem. Inf. Model.* **2010**, *50*, 1005−1011.
(31) Rauhamäki, S.; Postila, P. A.; Lätti, S.; Niinivehmas, S.; Multamäki, E.; Liedl, K. R.; Pentikäinen, O. T. Discovery of Retinoic Acid-Related Orphan Receptor Γt Inverse Agonists via Docking and Negative Image-Based Screening. *ACS Omega* **2018**, *3*, 6259−6266.
(32) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49*, 492−502. Software can be found at http://users.abo.fi/mivainio/shaep/ (accessed 02/19/2018).
(33) Egea, P. F.; Moras, D.; Biologie, L. De Molecular Recognition of Agonist Ligands by RXRs. *Mol. Endocrinol.* **2002**, *16*, 987−997.
(34) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461. Software can be found at http://vina.scripps.edu/download.html (accessed 04/28/2018).
(35) Lehtonen, J. V.; Still, D.-J.; Rantanen, V.-V.; Ekholm, J.; Björklund, D.; Iftikhar, Z.; Huhtala, M.; Repo, S.; Jussila, A.; Jaakkola, J.; Pentikäinen, O.; Nyrönen, T.; Salminen, T.; Gyllenberg, M.; Johnson, M. S. BODIL: A Molecular Modeling Environment for Structure-Function Analysis and Drug Design. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 401−419. Software can be found at http://users.abo.fi/bodil/about.php (accessed 04/21/2005).
(36) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38. Software can be found at http://www.ks.uiuc.edu/Research/vmd/ (accessed 02/19/2018).
(37) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96. Software can be found at http://www.mnf.uni-tuebingen.de/fachbereiche/pharmazie-und-biochemie/pharmazie/pharmazeutische-chemie/pd-dr-t-exner/research/plants.html (accessed 05/09/2016).
(38) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(39) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(40) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(41) 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.
(42) 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.;

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring 1. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(43) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132−1156. Software can be found at http://dock.compbio.ucsf.edu/Online_Licensing/index.html (accessed 04/27/2017).
(44) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791. Software can be found at http://autodock.scripps.edu/downloads/autodock-registration/autodock-4-2-download-page/ (accessed 04/28/2018).
(45) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893−1904. Software can be found at https://sourceforge.net/projects/smina/files/ (accessed 04/09/2018).
(46) Wang, J. L.; Limburg, D.; Graneto, M. J.; Springer, J.; Hamper, J. R. B.; Liao, S.; Pawlitz, J. L.; Kurumbail, R. G.; Maziasz, T.; Talley, J. J.; Kiefer, J. R.; Carter, J. The Novel Benzopyran Class of Selective Cyclooxygenase-2 Inhibitors. Part 2: The Second Clinical Candidate Having a Shorter and Favorable Human Half-Life. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 7159−7163.
(47) Card, G. L.; England, B. P.; Suzuki, Y.; Fong, D.; Powell, B.; Lee, B.; Luu, C.; Tabrizizad, M.; Gillette, S.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S.-H.; Schlessinger, J.; Zhang, K. Y. J. Structural Basis for the Activity of Drugs That Inhibit Phosphodiesterases. *Structure* **2004**, *12*, 2233−2247.
(48) Sung, B.-J.; Yeon Hwang, K.; Ho Jeon, Y.; Lee, J. I.; Heo, Y.-S.; Hwan Kim, J.; Moon, J.; Min Yoon, J.; Hyun, Y.-L.; Kim, E.; Jin Eum, S.; Park, S.-Y.; Lee, J.-O.; Gyu Lee, T.; Ro, S.; Myung Cho, J. Structure of the Catalytic Domain of Human Phosphodiesterase 5 with Bound Drug Molecules. *Nature* **2003**, *425*, 98−102.
(49) Bledsoe, R. K.; Madauss, K. P.; Holt, J. A.; Apolito, C. J.; Lambert, M. H.; Pearce, K. H.; Stanley, T. B.; Stewart, E. L.; Trump, R. P.; Willson, T. H.; Williams, S. P. A Ligand-Mediated Hydrogen Bond Network Required for the Activation of the Mineralocorticoid Receptor. *J. Biol. Chem.* **2005**, *280*, 31283−31293.
(50) Finley, J. B.; Atigadda, V. R.; Duarte, F.; Zhao, J. J.; Brouillette, W. J.; Air, G. M.; Luo, M. Novel Aromatic Inhibitors of Influenza Virus Neuraminidase Make Selective Interactions with Conserved Residues and Water Molecules in the Active Site. *J. Mol. Biol.* **1999**, *293*, 1107−1119.
(51) Kranjc, A.; Bongarzone, S.; Rossetti, G.; Biarnés, X.; Cavalli, A.; Bolognesi, M. L.; Roberti, M.; Legname, G.; Carloni, P. Docking Ligands on Protein Surfaces: The Case Study of Prion Protein. *J. Chem. Theory Comput.* **2009**, *5*, 2565−2573.
(52) Lätti, S.; Niinivehmas, S.; Pentikäinen, O. T. Rocker: Open Source, Easy-to-Use Tool for AUC and Enrichment Calculations and ROC Visualization. *J. Cheminf.* **2016**, *8*, 1−5. Software can be found at http://www.medchem.fi/rocker/ (accessed 02/19/2018).
(53) Halgren, T. A. Potential Energy Functions. *Curr. Opin. Struct. Biol.* **1995**, *5*, 205−210.
(54) Ahinko, M.; Kurkinen, S. T.; Niinivehmas, S. P.; Pentikäinen, O. T.; Postila, P. A. A Practical Perspective: The Effect of Ligand Conformers on the Negative Image-Based Screening. *Int. J. Mol. Sci.* **2019**, *20*, 2779.
(55) Bauer, M. R.; Mackey, M. D. Electrostatic Complementarity as a Fast and Effective Tool to Optimize Binding and Selectivity of Protein-Ligand Complexes. *J. Med. Chem.* **2019**, *62*, 3036−3050.
(56) Chaput, L.; Martinez-Sanz, J.; Quiniou, E.; Rigolet, P.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminf.* **2016**, *8*, 1−17.
(57) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Sscoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.
(58) Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.
(59) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(60) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(61) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.
(62) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1−20.
(63) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R Grand Challenge 2015: Evaluation of Protein-Ligand Pose and Affinity Predictions. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 651−668.
(64) Kelley, B. P.; Brown, S. P.; Warren, G. L.; Muchmore, S. W. POSIT: Flexible Shape-Guided Docking for Pose Prediction. *J. Chem. Inf. Model.* **2015**, *55*, 1771−1780.
(65) Kumar, A.; Zhang, K. Y. J. Application of Shape Similarity in Pose Selection and Virtual Screening in CSARdock2014 Exercise. *J. Chem. Inf. Model.* **2016**, *56*, 965−973.
(66) Gao, C.; Thorsteinson, N.; Watson, I.; Wang, J.; Vieth, M. Knowledge-Based Strategy to Improve Ligand Pose Prediction Accuracy for Lead Optimization. *J. Chem. Inf. Model.* **2015**, *55*, 1460−1468.
(67) Slynko, I.; Da Silva, F.; Bret, G.; Rognan, D. Docking Pose Selection by Interaction Pattern Graph Similarity: Application to the D3R Grand Challenge 2015. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 669−683.
(68) Kleywegt, G. J.; Jones, T. A. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1994**, *50*, 178−185.
(69) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.
(70) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **2012**, *8*, e1002708.
(71) Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A. POVME: An Algorithm for Measuring Binding-Pocket Volumes. *J. Mol. Graphics Modell.* **2011**, *29*, 773−776.
(72) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047−5056.
(73) Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J. Chem. Theory Comput.* **2017**, *13*, 4584−4592.
(74) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377−389.
(75) Halgren, T. New Method for Fast and Accurate Binding-Site Identification and Analysis. *Chem. Biol. Drug Des.* **2007**, *69*, 146−148.
(76) Totrov, M.; Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178−184.

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599

(77) B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing Protein Flexibility in Docking and Its Applications. *Drug Discovery Today* **2009**, *14*, 394−400.
(78) Ganesan, A.; Coote, M. L.; Barakat, K. Molecular Dynamics-Driven Drug Discovery: Leaping Forward with Confidence. *Drug Discovery Today* **2017**, *22*, 249−269.
(79) Postila, P. A.; Kaszuba, K.; Kuleta, P.; Vattulainen, I.; Sarewicz, M.; Osyczka, A.; Róg, T. Atomistic Determinants of Co-Enzyme Q Reduction at the Qi-Site of the Cytochrome Bc1 Complex. *Sci. Rep.* **2016**, *6*, 33607.
(80) Postila, P. A.; Kaszuba, K.; Sarewicz, M.; Osyczka, A.; Vattulainen, I.; Róg, T. Key Role of Water in Proton Transfer at the Qo-Site of the Cytochrome Bc1 Complex Predicted by Atomistic Molecular Dynamics Simulations. *Biochim. Biophys. Acta, Bioenerg.* **2013**, *1827*, 761−768.
(81) Postila, P. A.; Ylilauri, M.; Pentikäinen, O. T. Full and Partial Agonism of Ionotropic Glutamate Receptors Indicated by Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2011**, *51*, 1037−1047.
(82) Leanne Lash-Van Wyhe, L.; Postila, P. A.; Tsubone, K.; Sasaki, M.; Pentikäinen, O. T.; Sakai, R.; Swanson, G. T. Pharmacological Activity of C10-Substituted Analogs of the High-Affinity Kainate Receptor Agonist Dysiherbaine. *Neuropharmacology* **2010**, *58*, 640−649.
(83) Postila, P. A.; Swanson, G. T.; Pentikäinen, O. T. Exploring Kainate Receptor Pharmacology Using Molecular Dynamics Simulations. *Neuropharmacology* **2010**, *58*, 515−527.
(84) Frydenvang, K.; Lash, L. L.; Naur, P.; Postila, P. A.; Pickering, D. S.; Smith, C. M.; Gajhede, M.; Sasaki, M.; Sakai, R.; Pentikäinen, O. T.; Swanson, G. T.; Kastrup, J. S. Full Domain Closure of the Ligand-Binding Core of the Ionotropic Glutamate Receptor IGluR5 Induced by the High Affinity Agonist Dysiherbaine and the Functional Antagonist 8,9-Dideoxyneodysiherbaine. *J. Biol. Chem.* **2009**, *284*, 14219−14229.
(85) Lash, L. L.; Sanders, J. M.; Akiyama, N.; Shoji, M.; Postila, P.; Pentikäinen, O. T.; Sasaki, M.; Sakai, R.; Swanson, G. T. Novel Analogs and Stereoisomers of the Marine Toxin Neodysiherbaine with Specificity for Kainate Receptors. *J. Pharmacol. Exp. Ther.* **2007**, *324*, 484−496.
(86) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.
(87) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(88) Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.* **2017**, *1607*, 627−641.
(89) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(90) Gasteiger, J.; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, *19*, 3181−3184.
(91) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera-A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(92) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.
(93) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(94) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 1−14.
(95) Hanley, A. J.; McNeil, J. B. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29−36.

DOI: 10.1021/acs.jcim.9b00383
J. Chem. Inf. Model. 2019, 59, 3584−3599