Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy

Jason B. Cross,*,† David C. Thompson,⊥,‡ Brajesh K. Rai,#,§ J. Christian Baber,‡ Kristi Yi Fan,§ Yongbo Hu,|| and Christine Humblet§

Wyeth Research, Chemical Sciences, 500 Arcola Road, Collegeville, Pennsylvania 19426, Wyeth Research, Chemical Sciences, 200 Cambridge Park Drive, Cambridge, Massachusetts 02140, Wyeth Research, Chemical Sciences, 865 Ridge Road, Princeton, New Jersey 08543, and Wyeth Research, Chemical Sciences, 401 N. Middletown Road, Pearl River, New York 10965

Received February 13, 2009

Molecular docking programs are widely used modeling tools for predicting ligand binding modes and structure based virtual screening. In this study, six molecular docking programs (DOCK, FlexX, GLIDE, ICM, PhDOCK, and Surflex) were evaluated using metrics intended to assess docking pose and virtual screening accuracy. Cognate ligand docking to 68 diverse, high-resolution X-ray complexes revealed that ICM, GLIDE, and Surflex generated ligand poses close to the X-ray conformation more often than the other docking programs. GLIDE and Surflex also outperformed the other docking programs when used for virtual screening, based on mean ROC AUC and ROC enrichment values obtained for the 40 protein targets in the Directory of Useful Decoys (DUD). Further analysis uncovered general trends in accuracy that are specific for particular protein families. Modifying basic parameters in the software was shown to have a significant effect on docking and virtual screening results, suggesting that expert knowledge is critical for optimizing the accuracy of these methods.

INTRODUCTION

Molecular docking programs are commonly used to position small molecules within a three-dimensional representation of the protein structure. As the number of protein X-ray structures has increased dramatically in recent years,[1] these programs have become standard computational tools used in structure-based optimization of lead compounds.[2] Increasingly, molecular docking programs are being used to find novel ligands through virtual screening of compound libraries.[3]

There are a large number of docking programs currently available, with new programs constantly being developed and many existing programs continually being upgraded with new technology. Due to this constantly shifting landscape, publications comparing the performance, strengths, and deficiencies of these programs can be of real value to the chemistry community. In recent years, there have been a number of evaluations that have focused primarily on pose prediction,[4−7] virtual screening,[8−18] or a combination of these capabilities.[19−23] However, combining the results of these evaluations to form a coherent assessment of a docking program can be difficult, since unique data sets and different data analysis methods are often employed.

Several recent publications[24−28] have stressed the need for standards in the experimental design and reporting of results from docking evaluations. Data set composition and preparation can strongly influence reported success rates due to the diversity of targets and ligands studied as well as minor differences in methodology.[24,28−30] To this end, use of publicly available, standardized data sets like the Directory of Useful Decoys (DUD)[31,32] can be helpful. There are also a growing number of metrics being used to report success rates in pose prediction and virtual screening, all with specific strengths and weaknesses.[28,33] Many docking evaluation studies also lack the rigorous statistical analysis that should be a prerequisite for claiming real performance differences between docking programs.[24,26]

In this study, several docking programs (DOCK, FlexX, GLIDE, ICM, PhDOCK, and Surflex) have been evaluated by comparing their ability to generate and identify docking poses that are close to the X-ray conformation (cognate ligand docking) and their utility in selecting active compounds from a database of decoys (virtual screening). The docking programs were selected based primarily on availability to the authors; the software was either already in use or developed in-house, or demo versions had been arranged with commercial vendors. Although there are many additional docking programs available, it was not the authors' intent to survey all those currently in widespread use, and many of those examined are in regular use throughout the pharmaceutical industry. Programs were evaluated using their default settings, which provided baseline performance in assuming no prior experience on the part of the user with the target or software. This eliminated user subjectivity that could bias the results. Statistical analyses have been applied

* Corresponding author e-mail: jason.cross@cubist.com; Current address: Cubist Pharmaceuticals, Inc., 65 Hayden Ave., Lexington, MA 02421.
† Wyeth Research, Chemical Sciences, Collegeville, PA.
⊥ Current address: Boehringer Ingelheim Pharmaceuticals Inc., 900 Ridgebury Road, Ridgefield, CT 06877.
‡ Wyeth Research, Chemical Sciences, Cambridge, MA.
# Current address: Pfizer Global Research & Development, 260 Eastern Point Road, Groton, CT 06340.
§ Wyeth Research, Chemical Sciences, Princeton, NJ.
|| Wyeth Research, Chemical Sciences, Pearl River, NY.

10.1021/ci900056c CCC: $40.75
Published on Web 05/28/2009

to the docking and virtual screening results in an attempt to meaningfully differentiate the performance of these programs.

COMPUTATIONAL METHODS

Cognate Ligand Docking. The data set of X-ray complexes consisted of a subset of the CCDC/Astex test set[34] of publicly available structures from the Protein Data Bank (PDB)[35] along with additional complexes chosen due to pharmaceutical interest. Structures from the CCDC/Astex test set were selected based on resolution (<2.0 Å). Furthermore, complexes containing a heme group or covalent interactions between protein and ligand were removed. Ten kinase complexes (9 from the PDB and one in-house) and 12 nuclear receptor complexes (9 from the PDB and three in-house) were also included in the data set, since the high resolution CCDC/Astex test set contained few examples of these proteins. In addition, complexes were eliminated if the version of CORINA[36] used in this study did not produce a topologically correct chemical structure for the ligand from a SMILES string. A list of PDB accession codes and details of the X-ray complexes is given in Table 1.

X-ray costructures were downloaded from the PDB.[35] Ligands, water molecules, and counterions were excised from each structure. Visual inspection of amino acids resulted in correction of asparagine and gluatamine side chain conformations and histidine protonation states based on hydrogen bonding networks present in the protein. Hydrogen atom addition and minimization was performed using the Protein Preparation Wizard in Maestro[37] in the absence of the ligand. All heavy atoms remained fixed in their X-ray positions and were not minimized.

The X-ray conformation of each ligand was visually inspected, bond orders were assigned, and hydrogen atoms were added using the Maestro interface. Low energy ligand 3D conformations were also generated from SMILES strings using CORINA in the absence of the protein, since the docking programs studied generally required a 3D ligand conformation as input. CORINA was set to generate all possible stereoisomers as well as a maximum of 10 ring conformations.

Virtual Screening. The Directory of Useful Decoys (DUD)[31] was used to evaluate the performance of docking programs in virtual screening. The DUD data set consists of 40 diverse sets of protein targets, active and decoy compounds. Decoys were chosen based on similarity of their physical properties compared to the actives while still being topologically distinct, making this a challenging test set. All structures were downloaded from the DUD Web site (http://dud.docking.org). Although there are known issues and limitations within the DUD, it remains a valuable virtual screening benchmark available to the entire chemistry community.[32] Protein preparation was carried out using the Protein Preparation Wizard in Maestro.[37] Hydrogen atom addition and minimization was carried out in the absence of the cognate ligand. Protein heavy atoms were not minimized and retained their X-ray structure coordinates. Protonation states for histidine and side chain conformations for asparagine and glutamine were determined by visual inspection of the protein. Active and decoy ligand sets were used as provided from the DUD Web site with no additional computational manipulation.

Table 1. 68 X-ray Complexes Used for Cognate Ligand Docking

PDB code	resolution (Å)	protein
1A28	1.80	progesterone receptor
1A4Q	1.90	neuraminidase
1A6W	2.00	B1−8 FV Ab fragment
1ABE	1.70	L-arabinose-binding protein
1ABF	1.90	L-arabinose-binding protein
1AOE	1.60	dihydrofolate reductase
1AQW	1.80	glutathione S-transferase
1ATL	1.80	atrolysin C
1BMA	1.80	elastase
1C83	1.80	protein tyrosine phosphatase 1B
1COY	1.80	cholesterol oxidase
1D3H	1.80	dihydroorotate dehydrogenase
1E8W	2.50	phosphoinositide 3-kinase
1F3D	1.87	catalytic Ab 4B2
1FCZ	1.38	retinoic acid receptor
1FLR	1.85	4−4−20 FAb fragment
1FM9	2.10	retinoid X receptor
1GLQ	1.80	glutathione S-transferase
1HFC	1.50	collagenase
1HSL	1.89	histidine-binding protein
1HVR	1.80	HIV protease
1HYT	1.70	thermolysin
1JAP	1.82	MMP8
1JPA	1.91	Ephb2 receptor tyrosine kinase
1KE5	2.20	CDK2
1L2I	1.95	estrogen receptor
1LNA	1.90	thermolysin
1LST	1.80	LAO-binding protein
1MLD	1.83	malate dehydrogenase
1MMQ	1.90	matrilysin
1MRG	1.80	momorcharin
1MRK	1.60	trichosanthin
1MTS	1.90	Factor Xa
1MVC	1.90	retinoid X receptor
1NCO	1.80	neocarzinostatin
1NHZ	2.30	glucocorticoid receptor
1NQ7	1.50	retinoid-related orphan receptor
1O61	1.90	PLP-dependent enzyme
1QBR	1.80	HIV protease
1QBU	1.80	HIV protease
1QPC	1.60	leukocyte-specific protein tyrosine kinase
1QPD	2.00	leukocyte-specific protein tyrosine kinase
1SLT	1.90	S-lectin
1SRJ	1.80	streptavidin
1TMN	1.90	thermolysin
1TXI	1.90	vitamin D receptor
1TYL	1.90	insulin
1WAP	1.80	Trp RNA-binding attenuation protein
1XID	1.70	D-xylose isomerase
1XIE	1.70	D-xylose isomerase
1YDR	2.20	cAMP-dependent protein kinase
1Z95	1.80	androgen receptor
2AA2	1.95	mineralocorticoid receptor
2CL1	2.20	Nek2 centrosomal kinase
2CMD	1.87	malate dehydrogenase
2CTC	1.40	carboxypeptidase A
2GBP	1.90	galactose chemoreceptor protein
2QWK	1.80	neuraminidase
2SRC	1.50	C-src tyrosine kinase
2TMN	1.60	thermolysin
3ERT	1.90	estrogen receptor
3TPI	1.90	trypsinogen
4DFR	1.70	dihydrofolate reductase
5ABP	1.80	L-arabinose-binding protein
in-house #1	2.33	kinase
in-house #2	2.00	nuclear hormone receptor
in-house #3	1.93	nuclear hormone receptor
in-house #4	2.15	nuclear hormone receptor

Docking Programs. Specific details on how each docking program was run are given in the following sections. The versions of the docking programs included in this study were the most up to date available at the time of the evaluation, although several newer versions have been released since then. All docking programs were used in their default configuration with no tuning of optional parameters, unless otherwise noted.

DOCK 6.1. The flexible anchor-and-grow algorithm[38,39] available in DOCK 6.1 was used in this study. Parameters were the same as in a previous study.[40] The DMS program, distributed with DOCK 6.1, was used to generate a molecular surface for each receptor.[41] The SPHGEN utility was then used to create the negative image of the surface,[42,43] and the sphere set for each complex was composed of all spheres found within 10 Å of any ligand atom. A receptor box was generated using the SHOWBOX utility and was centered on the ligand with an additional 5 Å boundary. The GRID utility was used to precalculate the scoring function potential grids.

FlexX (v2.0.3). Docking was carried out using the standalone FlexX[44] package. Default parameters for base fragment placement and incremental construction were used for docking. The FlexX scoring function, adapted from work by Boehm[45,46] and Klebe,[47] was used for pose scoring. Formal charges were used throughout.

GLIDE (v4.5). Docking using the GLIDE algorithm[48,49] utilizes precomputed grids and occurs in a hierarchical fashion. Grids were generated using a receptor site that was defined by the centroid of the cognate ligand. The docking hierarchy starts with the systematic conformational expansion of the ligand, followed by placement in the receptor site. Minimization of the ligand in the field of the receptor is then carried out using the OPLS-AA[50] force field with a distance-dependent dielectric (default = 2.0). The lowest energy poses are then subjected to a Monte Carlo procedure that samples nearby torsional minima. The best pose for a given ligand is determined by the composite Emodel score. Different compounds can then be ranked using GlideScore, a modified version of the ChemScore[51] function that includes terms for steric clashes and buried polar groups. Default van der Waals scaling was used (1.0 for the receptor and 0.8 for the ligand).

GLIDE v4.5 has a set of three choices for default docking behavior. Standard-precision (SP) docking is the procedure described in the previous paragraph. High-throughput virtual screening (HTVS) docking is more suitable for rapid screening of large ligand databases. Conformational sampling is significantly reduced in HTVS compared to SP, increasing the computational speed. Extra-precision (XP) docking[52] is designed to reduce the false positive rate. Sampling is more extensive, using the results from SP docking as a starting point for a high-resolution anchor-and-grow strategy. The XP scoring function contains a number of additional terms beyond those present in GlideScore, including terms for hydrophobic enclosure and large desolvation penalties. In this study, the SP and XP protocols for cognate ligand docking and the HTVS protocol for virtual screening have been used. The SP protocol was not included in the main section of the virtual screening study, since this method is intended for pose prediction based on statements in the user manual, but was included as an example of how tuned parameters can have a significant effect on virtual screening results due to its known virtual screening capabilities.[49]

ICM (v3.5−1). The ICM methodology[53,54] utilizes internal coordinates to optimize flexible ligands in a grid-based receptor field. The grid potentials include electrostatic, hydrogen bond, hydrophobic, and two van der Waals terms. Energies are computed using MMFF[55] partial charges with the ECEPP/3 force field.[56] Global optimization starts with a random conformational change of the free bonds, angles, and torsions according to the Biased Probability Monte Carlo[57] (BPMC) algorithm followed by local energy minimization of the analytical differentiable terms. Docking was performed against a rigid receptor.

PhDOCK. The PhDOCK approach[58,59] has been investigated, as implemented using DOCK4.0.[60] During the docking, each database ensemble (a 3D pharmacophore and associated conformers) is simultaneously docked into the target binding site and then scored using the contact scoring function. The contact function is a summation of the number of heavy atom contacts arising between the ligand and receptor atoms up to 4 Å away.[38] Ten poses per molecule from PhDOCK are then rescored using a high-throughput solvation based MM-PBSA scoring function to approximate the free energy of binding.[61] This workflow has been developed in-house and is intended for use in virtual high-throughput screening.

A PhDOCK database consists of conformers of the same or different molecules overlaid by their largest three-dimensional (3D) pharmacophore. MCSS2SPTS[62] was used to determine the site points that are used for the matching during the docking. Theoretical pharmacophore points, representing hot spots in the binding site, allow for preferential orientation of the ligands in productive modes. For this study, a conformationally expanded PhDOCK database was generated for each test set ligand. For the cognate ligand docking portion of the study, the translated bound conformation of each ligand was also included in the database. Starting from this bound conformation of the ligand, a maximum of 1000 conformers were generated using OMEGA2[63] with a 25 kcal/mol internal energy cutoff evaluated with the MMFF94s force field.[64]

During the rescoring of docked poses, for each complex, the ligand was subjected to up to 1000 steps of Cartesian coordinate minimization within the fixed protein structure using the Szybki minimizer[65] and the MMFF94s force field;[64] the default convergence criteria of 0.1 kcal/mol·Å on the gradient vector norm was used: $\sum_i \sqrt{g_i g_i}$. Each of the terms comprising the binding free energy were approximated by calculating the value for the complex, the protein alone, and the ligand alone, where the structures for the protein and ligand are simply taken from the minimized complex. Poisson−Boltzmann electrostatics were computed using ZAP[66] with Bondi radii[67] to include solvation effects. Additional van der Waals energy terms were calculated within Szybki. Default parameters were used.[68]

Surflex (v2.1). The Surflex algorithm[69−71] employs the empirical Hammerhead[72] scoring function and uses an idealized active site ligand, or protomol, to generate ligand poses by incremental construction and a crossover procedure that combines pieces from distinct poses. Protomols were computed using the position of the cognate ligand to define the binding site. Docking was performed with the executables

Table 2. Physical Properties of the 68 Ligands Used for Cognate Docking

property	mean	std. dev.	minimum	maximum
		Entire Data Set		
molecular weight	328.9	139.4	135.1	758.9
AlogP	0.5	3.7	−8.1	9.1
donors	3.0	2.0	0.0	8.0
acceptors	6.1	3.1	2.0	18.0
rotatable bonds	4.8	3.3	0.0	13.0
		Kinase Subset		
molecular weight	346.2	112.5	135.1	502.2
AlogP	−0.8	3.9	−8.1	5.0
donors	2.8	1.6	1.0	6.0
acceptors	8.0	4.6	5.0	18.0
rotatable bonds	3.3	2.4	0.0	8.0
		Nuclear Receptor Subset		
molecular weight	411.7	87.0	320.4	617.6
AlogP	4.8	2.0	1.2	9.1
donors	1.2	0.9	0.0	3.0
acceptors	4.1	1.9	2.0	7.0
rotatable bonds	5.3	3.2	2.0	12.0

distributed by Tripos, Inc. Docking studies were conducted with the ring flexibility option turned off (default) and turned on. While the addition of ring flexibility strictly meets the definition of a tuned parameter in the context of this study, the authors suggest that this setting more closely reflects how Surflex would be used "out of the box" even by a novice user. Docking studies were repeated with the ring flexibility option in order to test the effect of parameter tuning on docking and virtual screening performance.

RESULTS AND DISCUSSION

Cognate Ligand Docking Accuracy. The ability of a docking program to reproduce a ligand pose close to that found in an X-ray complex is often a critical determinant of the program's effectiveness for structure-guided design. In this evaluation, a data set (Table 1) of X-ray complexes that contained ligands with diverse physical properties was used. Pipeline Pilot was used to calculate these properties, which are listed in Table 2. Each of the cognate ligands was docked back into its respective protein structure using different starting conformations as input. In the first case, the initial ligand conformation was taken directly from the X-ray complex. In the second case, a low energy conformation of the ligand generated by CORINA was used as input for docking, the aim being to remove any conformational and positional information from the ligand before docking. For each of the ligand conformations used as docking input (X-ray and CORINA) the 10 top scoring poses were retained for further analysis. If fewer than 10 docked poses were generated for a particular complex, all poses were used in the data analysis. The heavy atom root-mean-squared deviation (RMSD) from the X-ray ligand conformation was computed for each docked pose using an internally developed OEChem[73] script. Molecular symmetry was accounted for in this algorithm, eliminating artificially high RMSDs due to issues such as phenyl ring flipping. RMSDs were tabulated for both the top scoring pose and for the lowest-RMSD (best) pose generated by the docking program.

Statistical averages, standard deviations, and 95% confidence intervals for the RMSD values of the top scoring and best poses are listed in Table 3. Confidence intervals were computed using bootstrapping techniques, where the mean was calculated 10,000 times with 20% of the RMSD data removed each iteration. Table 4 lists the p-values (calculated using the paired *t* test) and Pearson correlations for the mean RMSDs using the top scoring pose for each pair of docking programs, while the mean differences and correlation corrected 95% confidences are listed in the Supporting Information. Statistical significance can be assessed based on a p-value <0.05 or by comparing the mean difference and the correlation corrected 95% confidence; if the lower confidence limit is greater than 0.0, the result is statistically significant. The mean differences did approximate the normal distribution, making the use of the paired *t* test appropriate in this

Table 3. Statistical Results of Cognate Ligand Docking for the Data Set of 68 X-ray Complexes

docking program	median (Å)	mean (Å)	std. dev. (Å)	95% CI[a] (Å)	no. of ligands[b]
		Input Ligand Conformation from X-ray Complex			
		Top Scoring Pose			
DOCK	0.95	2.62	3.14	2.22−2.97	67
FlexX	1.79	3.32	3.61	2.85−3.69	66
GLIDE SP	0.81	1.38	1.57	1.18−1.54	67
GLIDE XP	0.62	1.16	1.50	0.97−1.30	67
ICM	0.79	1.73	2.12	1.46−1.96	68
PhDOCK	0.91	2.56	3.24	2.15−2.89	67
Surflex	1.19	1.97	1.94	1.73−2.19	68
		Tuned Parameters			
GLIDE HTVS	2.12	2.39	2.17	2.10−2.65	63
Surflex Ringflex	1.05	2.07	2.09	1.80−2.30	68
		Best Pose			
DOCK	0.88	1.88	2.43	1.58−2.13	67
FlexX	0.80	2.06	2.82	1.68−2.32	66
GLIDE SP	0.52	0.88	1.07	0.74−0.97	67
GLIDE XP	0.55	0.81	0.84	0.70−0.89	67
ICM	0.60	1.02	1.22	0.85−1.13	68
PhDOCK	0.72	1.85	2.25	1.56−2.08	67
Surflex	0.74	1.48	1.54	1.28−1.65	68
		Tuned Parameters			
GLIDE HTVS	0.88	1.63	1.75	1.39−1.83	63
Surflex Ringflex	0.70	1.52	1.79	1.28−1.70	68
		Input Ligand Conformation from CORINA			
		Top Scoring Pose			
DOCK	1.32	3.28	3.53	2.83−3.66	67
FlexX	3.09	4.27	4.63	3.67−4.78	64
GLIDE SP	1.22	2.08	2.49	1.75−2.35	65
GLIDE XP	1.15	1.97	2.13	1.70−2.20	66
ICM	0.98	1.86	1.97	1.61−2.08	68
PhDOCK	3.70	4.57	3.97	4.08−5.01	67
Surflex	1.66	2.93	2.57	2.62−3.24	68
Tuned Parameters					
GLIDE HTVS	2.12	2.66	2.29	2.37−2.92	62
Surflex Ringflex	1.19	2.26	2.30	1.97−2.51	68
		Best Pose			
DOCK	1.12	2.31	2.92	1.93−2.61	67
FlexX	2.13	3.33	4.09	2.78−3.74	64
GLIDE SP	0.72	1.30	1.40	1.11−1.45	65
GLIDE XP	0.79	1.34	1.53	1.14−1.50	66
ICM	0.71	1.23	1.24	1.07−1.37	68
PhDOCK	2.81	3.57	3.31	3.16−3.93	67
Surflex	1.27	2.06	1.94	1.81−2.28	68
		Tuned Parameters			
GLIDE HTVS	1.28	1.83	1.85	1.59−2.03	62
Surflex Ringflex	0.90	1.50	1.42	1.32−1.67	68

[a] 95% confidence interval. [b] Number of ligands successfully docked.

Table 4. Analysis of the Mean RMSDs for the Top Scoring Poses from Cognate Ligand Docking[a]

	DOCK	FlexX	GLIDE SP	GLIDE XP	ICM	PhDOCK	Surflex	GLIDE HTVS	Surflex Ringflex
Input Ligand Conformation from X-ray Complex									
DOCK		0.37	0.17	0.22	0.27	0.46	0.43	0.34	0.37
FlexX	**0.048**		0.15	0.18	0.18	0.36	0.44	0.09	0.18
GLIDE SP	**2.7E-3**	**7.2E-5**		0.76	0.17	0.13	0.19	0.31	0.07
GLIDE XP	**3.4E-4**	**9.0E-6**	0.96		0.12	0.25	0.28	0.25	0.17
ICM	**0.032**	**8.7E-4**	0.22	0.054		0.21	0.40	0.06	0.03
PhDOCK	0.89	0.10	**5.7E-3**	**5.9E-4**	0.056		0.33	0.23	0.11
Surflex	0.076	**6.9E-4**	**0.030**	**1.9E-3**	0.38	0.14		-0.05	0.36
GLIDE HTVS	0.55	0.11	**1.5E-4**	**1.4E-4**	0.092	0.74	0.19		-0.03
Surflex Ringflex	0.13	**9.7E-3**	0.25	**2.1E-3**	0.35	0.25	0.73	0.54	
Input Ligand Conformation from CORINA									
DOCK		0.27	0.41	0.38	0.42	0.25	0.46	0.48	0.48
FlexX	0.070		0.20	0.34	0.12	0.37	0.22	0.41	0.15
GLIDE SP	**0.010**	**1.1E-3**		0.62	0.63	0.48	0.55	0.52	0.72
GLIDE XP	**4.1E-3**	**1.5E-4**	0.67		0.49	0.40	0.39	0.65	0.28
ICM	**6.5E-4**	**1.4E-4**	0.32	0.59		0.21	0.39	0.48	0.47
PhDOCK	**0.025**	0.75	**6.5E-7**	**3.2E-7**	**7.4E-7**		0.44	0.36	0.37
Surflex	0.42	**0.025**	**8.9E-3**	**4.3E-3**	**9.8E-4**	**6.1E-4**		0.29	0.67
GLIDE HTVS	0.15	**9.3E-3**	**0.011**	**1.6E-3**	**0.011**	**2.9E-3**	0.71		0.21
Surflex Ringflex	**0.011**	**1.3E-3**	0.54	0.43	0.15	**5.2E-6**	**6.6E-3**	0.16	

[a] The upper triangles contain the Pearson correlations and the lower triangles contain the p-values (calculated using the paired t-test). p-Values in bold are statistically significant (95%). High Pearson correlations (> 0.80) are bold. Examples of modified parameter settings are shaded.

circumstance. Correlation of the RMSDs simply gives a measure of whether two methods tended to succeed and fail for the same complexes, not whether the same types of poses were generated. Overall, the docking programs were consistently better at generating and identifying low-RMSD docking poses when the X-ray conformation was used as input rather than when a low energy CORINA conformation was used. This result is unsurprising at first glance and has been noted previously,[23] since the X-ray conformation provides a best case scenario of supplying the "ideal" solution as input. Based on the current data and published methodologies, it is clear that none of the docking programs simply

starts with the input geometry of the ligand. If this were the case, the results for docking the X-ray input conformation would likely be much better than those observed.

The arithmetic mean RMSDs for the top scoring poses using the X-ray ligand conformation ranged from 1.16 to 3.32 Å across all docking programs, while the median RMSDs ranged from 0.62 to 1.79 Å. The medians tended to be lower than means due to the fact that the mean RMSD can be skewed by a few very high RMSD values and the lower limit of RMSD is bounded at 0 Å. Docking failures were also accounted for in the median RMSD by giving them an arbitrarily high RMSD value, while they were not included in the mean RMSD. GLIDE, both SP and XP versions, and ICM had the lowest average (mean and median) RMSD values of the programs evaluated, although many of the averages for the other docking programs were also below 2.0 Å. In fact, the mean RMSD for GLIDE (SP and XP) showed statistical significance (p-value ≤ 0.05) when compared to all other docking programs except ICM. The mean RMSD of ICM was significantly lower than DOCK and FlexX, while Surflex performed better than FlexX. Other differences between docking programs were not statistically significant based on mean RMSD. Correlation of the RMSDs was generally low to moderate, reaching a high of 0.76 when comparing GLIDE SP and GLIDE XP. Looking at the best docking poses generated using the X-ray ligand conformation, the mean RMSDs ranged from 0.81 to 2.06 Å and the median RMSDs ranged from 0.52 to 0.88 Å. GLIDE SP and XP as well as ICM yielded the lowest average RMSDs in this case as well. However, it must be noted that all docking programs in this study performed admirably in generating low RMSD poses, with only FlexX having mean RMSD above 2.0 Å.

The examination of the top scoring poses using the CORINA generated initial ligand conformations provided the most difficult test for reproducing the X-ray complex in this study, since the crystallographically determined coordinates were not supplied to the docking programs as input. This procedure also mimics the "real world" situation in which the optimal docking pose is not known *a priori*. Hence, these results may prove more instructive in assessing the docking programs for accuracy in reproducing the X-ray ligand pose. However, one must keep in mind that an additional variable has been introduced, namely the generation of a low energy ligand conformation by CORINA that is used as input for each docking program. The intention is that minimal bias has been established by using software that is not tied closely with any of the docking programs being examined.

The average RMSDs for the top scoring poses using the CORINA ligand conformations as input were notably higher than those seen when using the X-ray conformations, with mean values ranging from 1.86 to 4.57 Å and median values ranging from 0.98 to 3.70 Å. Although mean RMSDs tended to be quite high, with only ICM and GLIDE XP below 2.0 Å, median RMSDs were much lower overall, with only FlexX and PhDOCK above 2.0 Å. Again, the skewing of the mean RMSD was due to a relatively small number of complexes having very large RMSD values, while the RMSD value itself has a lower bound of 0.0 Å. ICM and GLIDE (XP and SP) produced the lowest mean RMSDs, which were significantly lower than DOCK, FlexX, PhDOCK, and Surflex. The mean RMSDs for Surflex and DOCK also showed varying levels of statistical significance when compared to FlexX and PhDOCK. The best docking poses for the CORINA ligand conformations were generally much better for all docking programs than the top scoring poses based on RMSD, though not as consistently low when compared to the best poses generated using the X-ray conformations. Mean RMSDs ranged from 1.23 to 3.57 Å and median RMSDs ranged from 0.71 to 2.81 Å. ICM had the lowest mean and median RMSD values compared to the other docking programs, slightly better than GLIDE SP and GLIDE XP. Most docking programs had reasonable median RMSDs, with only FlexX and PhDOCK above 2.0 Å.

Comparing the top scoring pose results to the best RMSD ones, it appears that while the docking programs studied are generally able to generate at least one pose with an RMSD below 2.0 Å, these low-RMSD poses are not consistently scored better than high-RMSD poses. This suggests that the scoring functions, in general, may be prone to errors in which the poses closest to the X-ray conformation do not consistently receive the best score, although a more detailed study of these scoring functions and their potential energy surfaces would be required for confirmation.[74] It also supports the view that when evaluating docking programs, it is necessary to look not only at the top scoring pose but also additional poses with worse scores. The cases in which no low-RMSD pose was found may be due to inadequacies in sampling rather than in the scoring function.[74]

As mentioned in the Computational Methods section, we intentionally seeded the data set with kinase and nuclear receptor X-ray complexes due to their interest as pharmaceutical drug targets. Docking results for these subsets, using the CORINA conformations as input, are summarized in Table 5. Even though there are relatively few complexes in each of these subsets, there were some noticeable trends when comparing the kinases and nuclear receptors to the average (mean and median) values for the entire data set. Another point of interest regarding the inclusions of these complexes is that they are not members of the highest resolution part of the CCDC/Astex data set, which has been used extensively for calibration and validation of docking programs. One would expect that docking programs that have not been tuned to this data set would show similar accuracy with these additional complexes, while those tools that have been could exhibit weakened performance.

For the kinase subset, average RMSD values for the top scoring poses were generally higher than or approximately equal to those for the composite data set. Due in part to the small sample size, statistical significance between the mean RMSD values was only seen for GLIDE XP, GLIDE SP, and Surflex when compared to PhDOCK (Table 6). However, when comparing the mean RMSDs of the best poses generated for the kinase subset the results were mixed, with some programs having higher mean RMSDs and some showing improved performance relative to the entire data set. Pearson correlation of the RMSDs was high between GLIDE SP and GLIDE XP as well as between Surflex and both GLIDE methodologies, although it must be noted that this metric only measures whether the programs succeeded or failed on the same complexes and not whether the same poses were generated. The physical properties of the kinase subset (Table 2) were only marginally different compared to the entire data set. The kinase subset has a lower mean

Table 5. Statistical Results of Cognate Ligand Docking for the Kinase and Nuclear Receptor Subsets Using the CORINA Ligand Conformations

docking program	median (Å)	mean (Å)	std. dev. (Å)	95% CI[a] (Å)	no. of ligands[b]
Kinase Subset					
Top Scoring Pose					
DOCK	3.28	3.86	3.69	2.56−4.62	10
FlexX	2.91	4.33	3.94	2.99−5.13	10
GLIDE SP	1.16	2.72	3.02	1.65−3.28	10
GLIDE XP	1.33	2.23	2.21	1.58−2.65	10
ICM	2.01	2.70	2.32	2.03−3.20	10
PhDOCK	4.63	4.72	1.72	4.21−5.19	10
Surflex	1.87	2.64	2.17	1.99−3.11	10
Tuned Parameters					
GLIDE HTVS	2.39	2.86	2.30	2.10−3.41	10
Surflex Ringflex	3.14	3.29	2.37	2.58−3.88	10
Best Pose					
DOCK	1.01	2.03	2.33	1.36−2.43	10
FlexX	2.23	2.33	1.27	1.97−2.69	10
GLIDE SP	1.14	1.72	1.78	1.26−2.05	10
GLIDE XP	0.82	2.06	2.22	1.39−2.44	10
ICM	0.73	1.37	1.34	0.92−1.59	10
PhDOCK	3.27	3.58	1.69	2.85−4.00	10
Surflex	1.15	1.68	1.80	1.05−1.98	10
Tuned Parameters					
GLIDE HTVS	1.93	2.26	1.88	1.69−2.68	10
Surflex Ringflex	1.33	1.29	0.73	1.09−1.48	10
Nuclear Receptor Subset					
Top Scoring Pose					
DOCK	0.88	2.39	3.83	1.44−2.77	12
FlexX	1.02	5.70	7.47	3.38−7.16	13
GLIDE SP	0.60	0.70	0.38	0.60−0.76	12
GLIDE XP	0.68	1.38	1.66	0.89−1.56	12
ICM	0.47	0.69	0.42	0.56−0.78	13
PhDOCK	0.95	4.41	5.75	2.73−5.23	12
Surflex	0.97	2.24	2.42	1.40−2.71	13
Tuned Parameters					
GLIDE HTVS	1.31	2.34	2.38	1.69−2.79	10
Surflex Ringflex	0.51	0.83	0.62	0.64−0.94	13
Best Pose					
DOCK	0.79	1.70	3.20	0.73−1.95	12
FlexX	0.96	5.04	7.05	2.71−6.33	13
GLIDE SP	0.44	0.61	0.37	0.51−0.69	12
GLIDE XP	0.53	0.70	0.32	0.61−0.76	12
ICM	0.47	0.64	0.35	0.54−0.73	13
PhDOCK	0.88	3.83	4.44	2.70−4.54	12
Surflex	0.93	1.55	1.82	0.97−1.82	13
Tuned Parameters					
GLIDE HTVS	0.60	1.43	1.97	0.85−1.70	10
Surflex Ringflex	0.51	0.80	0.61	0.61−0.90	13

[a] 95% confidence interval. [b] Number of ligands successfully docked.

AlogP, ∼2 more hydrogen bond acceptors and ∼1.5 fewer rotatable bonds on average compared to the entire data set. Again, these results suggest that poses with low RMSD compared to the X-ray conformation are not consistently scored better than poses with higher RMSDs, regardless of the docking program used, and this effect is more pronounced in the kinase subset compared to the whole data set of X-ray complexes. This effect could be due to scoring functions generally capturing hydrophobic interactions more accurately than hydrophilic ones, at least within the kinase data set, based on the increased hydrophilicity of the ligands.

Average RMSDs for the nuclear receptor subset (Table 5), for both the top scoring and best poses, were lower compared to the data set as a whole. The only exceptions to this were FlexX, which had a slightly higher mean RMSD (both for mean top and best RMSDs) yet an improved median RMSD compared to the entire data set, and PhDOCK, which had a higher mean RMSD for the best pose relative to the entire data set. ICM, GLIDE SP, and GLIDE XP all achieved statistical significance compared to FlexX, while other results were mixed (Table 6). Interestingly, the DOCK and GLIDE XP results were highly correlated, while the two GLIDE methodologies showed lower correlation. In fact, the generally more accurate GLIDE XP method did not perform as well as GLIDE SP. This difference in accuracy may be due to the flexible nature of the nuclear receptor binding sites as well as the ability of GLIDE SP to accommodate slight steric overlaps better than GLIDE XP. The physical properties (Table 2) of the nuclear receptor ligands differed more noticeably from the whole data set compared to the kinase ligands, with molecular weight ∼80 Da higher, AlogP more than 4 units higher, ∼2 fewer acceptors, and ∼2 fewer donors on average, and more closely resemble the property profile expected for known drugs. The difference in docking accuracy may be due to the properties of the nuclear receptor binding sites rather than the physical properties of the ligands, since non-nuclear receptor ligands in the data set that have similar property profiles did not generally yield low RMSD poses. It may also be due to training sets used in the development of docking programs that are weighted heavily toward druglike ligands. However, two of the kinases that have ligands with physical properties similar to the nuclear receptor ligands (1QPD and the in-house structure) did yield low RMSD poses with nearly all docking programs studied, suggesting that the improved performance may be due to a combination of ligand and receptor properties.

A common issue with this type of statistical approach for evaluating docking pose RMSDs became apparent when comparing the mean and median RMSD values for each of the docking programs. For example, DOCK had a respectable median RMSD of 1.32 Å for the top scoring poses using the CORINA conformations but had a mean RMSD value of 3.28 Å, which is much higher than the normally accepted cutoff of 2.0 Å for "good" docked poses. This suggested that a relatively small number of very poorly docked poses (i.e. poses with very high RMSDs) were primarily responsible for shifting the mean RMSD upward. It is also generally true that the distribution of RMSDs will be skewed, with the peak closer to the lower RMSD limit of 0 and a longer tail extending into higher RMSDs. This feature necessitated the use of bootstrapping to compute the 95% confidence intervals rather than using a method valid for normally distributed data.

A simple way to avoid this kind of bias is to plot the fraction of complexes that have poses below a given RMSD[25,28] (Figure 1). Using this reporting method it is straightforward to see how the RMSD values are distributed over a range of RMSDs, and it can be used to identify a "success rate" at an arbitrary RMSD cutoff. It also accounts for docking failures, which are difficult to include in the mean RMSD metric, by reducing the maximum success rate. Figure

Table 6. Analysis of the Mean RMSDs of the Top Scoring Poses from Cognate Ligand Docking for the Kinase and Nuclear Receptor Subsets Using the CORINA Ligand Conformations[a]

	DOCK	FlexX	GLIDE SP	GLIDE XP	ICM	PhDOCK	Surflex	GLIDE HTVS	Surflex Ringflex
Kinase Subset									
DOCK		-0.20	0.23	0.24	0.26	-0.13	0.24	**0.85**	0.11
FlexX	0.81		0.12	0.16	-0.08	0.55	0.21	-0.07	0.24
GLIDE SP	0.41	0.30		**0.97**	0.59	0.43	**0.89**	0.32	**0.85**
GLIDE XP	0.21	0.15	0.17		0.40	0.54	**0.86**	0.22	0.77
ICM	0.36	0.30	0.98	0.56		-0.23	0.23	0.51	0.77
PhDOCK	0.54	0.72	**0.047**	**2.9E-3**	0.076		0.39	-0.13	0.15
Surflex	0.34	0.22	0.89	0.29	0.95	**0.015**		0.25	0.69
GLIDE HTVS	0.17	0.35	0.89	0.49	0.83	0.087	0.80		0.24
Surflex Ringflex	0.68	0.44	0.28	0.057	0.26	0.13	0.27	0.64	
Nuclear Receptor Subset									
DOCK		0.50	0.01	**0.82**	0.04	0.18	0.40	**0.90**	-0.18
FlexX	0.080		0.20	0.38	0.26	0.34	0.01	**0.80**	0.33
GLIDE SP	0.14	**0.033**		0.34	0.42	0.79	-0.18	0.07	0.79
GLIDE XP	0.19	**0.046**	0.16		-0.24	0.42	0.42	0.64	0.19
ICM	0.14	**0.031**	0.90	0.23		0.13	-0.40	0.25	0.13
PhDOCK	0.29	0.47	0.067	0.12	**0.042**		-0.10	0.43	0.78
Surflex	0.93	0.14	**0.048**	0.17	0.053	0.29		0.26	-0.11
GLIDE HTVS	0.66	0.11	**0.042**	0.098	**0.048**	0.73	0.90		-0.15
Surflex Ringflex	0.19	**0.033**	0.20	0.30	0.48	**0.037**	0.071	0.056	

[a] The upper triangles contain the Pearson correlations, and the lower triangles contain the p-values (calculated using the paired t-test). p-Values in bold are statistically significant (95%). High Pearson correlations (> 0.80) are bold. Examples of modified parameter settings are shaded.

1 shows the top scoring and best pose RMSD results using the X-ray conformation as input. The relative rankings of the docking programs, using 2.0 Å as a cutoff, did not differ significantly compared to the mean and median RMSDs previously discussed, although the ranking for ICM did improve. At the 2.0 Å cutoff, ICM was able to correctly identify a top scoring pose below 2.0 Å for 77.9% of the complexes, while GLIDE XP and GLIDE SP were able to reach this threshold for 85.3% and 77.9% of the complexes, respectively. Figure 1 also shows the top scoring and best pose RMSD results when using the CORINA ligand conformations. In these cases, the results for ICM, GLIDE SP,

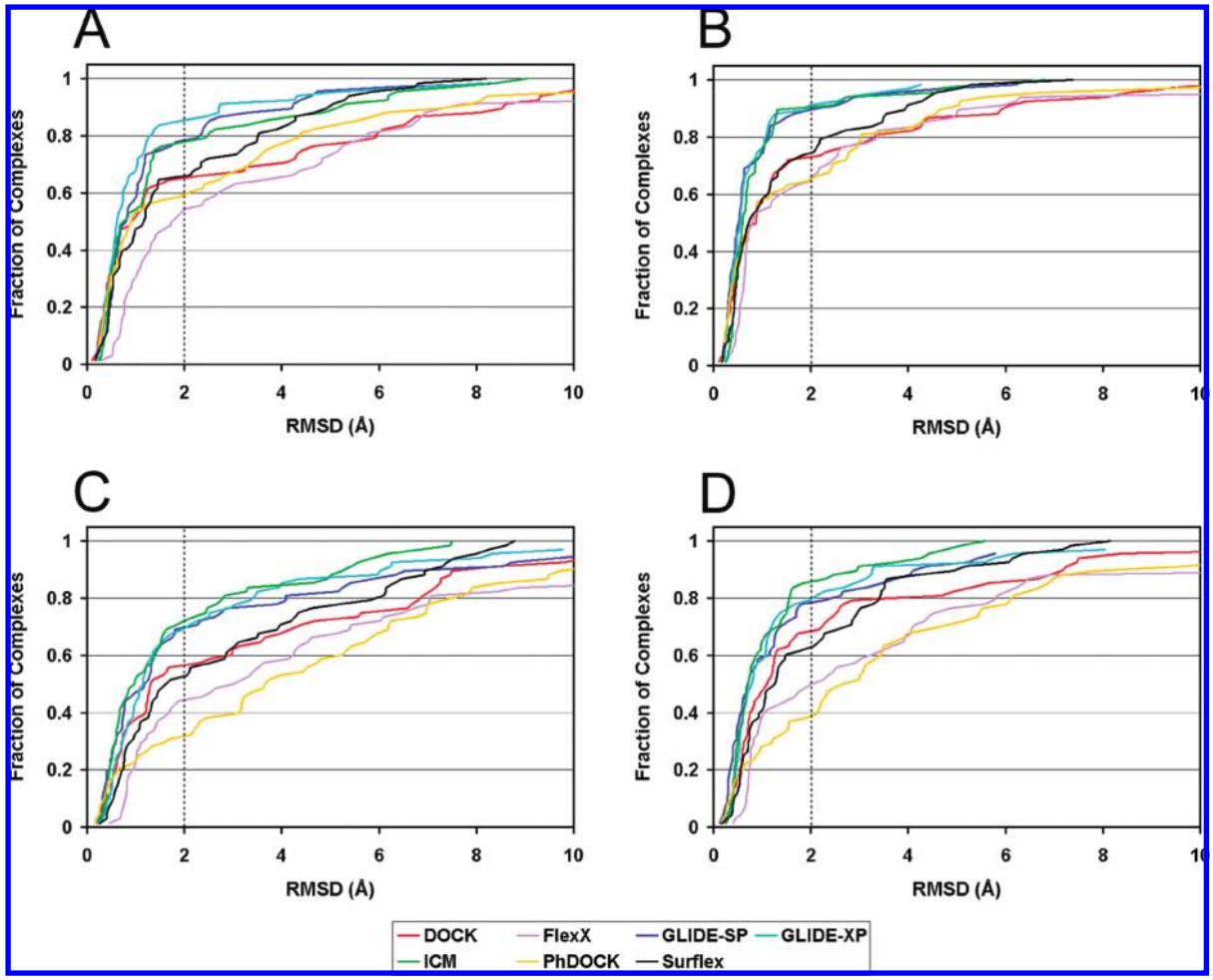

Figure 1. Cumulative distribution plots for cognate ligand docking of 68 protein−ligand complexes: (A) top scoring pose when the X-ray ligand conformation was used as input, (B) best pose when the X-ray ligand conformation was used as input, (C) top scoring pose when the CORINA ligand conformation was used as input, and (D) best pose when the CORINA ligand conformation was used as input. Dotted lines indicate a 2.0 Å RMSD cutoff.

and GLIDE XP using a 2.0 Å cutoff were essentially indistinguishable, with top scoring pose RMSDs less than 2.0 Å identified for 72.1%, 69.1%, and 69.1% of the complexes, respectively. It was also evident from the shapes of the lines in these graphs that the large mean RMSDs observed for some docking programs were the result of a relatively small proportion of poses that deviated significantly from the X-ray (i.e. top scoring CORINA poses for DOCK and Surflex), while in other cases it appeared that large deviation from the X-ray conformation was more widespread across the data set (i.e. top scoring CORINA poses for FlexX and PhDOCK).

Although this study was designed to compare docking programs using default parameters, each of these programs can be customized by changing a variety of settings. Experience working with a protein target often leads to a set of modified and well tested docking parameters that are tuned to the system of interest. This can lead to docking results that are far superior to those obtained with just the default software settings. As a brief example of how parameter tuning can affect docking results, the effects of some very simple parameter changes for the two docking programs that had the highest mean ROC AUCs for the DUD targets, GLIDE and Surflex, were investigated. (See the section on Virtual Screening Accuracy for details.) For GLIDE, the HTVS and SP methods were compared, since both use the same scoring function and only differ in the extent of conformational sampling employed; however, in this case GLIDE HTVS assumes the role of a tuned method, even though it is a faster and less accurate version of the program. For Surflex, the flexible ring switch (useful for saturated or partially saturated ring systems) was turned on during docking and has been designated Surflex Ringflex in the rest of this paper.

From the results listed in Table 3, it is clear that the average RMSDs for GLIDE HTVS are consistently larger than for GLIDE SP (mean and median), when the top scoring and best poses are considered. In fact, the difference in mean RMSD was statistically significant (Table 4) when both the X-ray and CORINA conformations were used for docking. This suggests that the additional conformational sampling present in the GLIDE SP method is critical for successfully generating low RMSD poses. The GLIDE HTVS average RMSDs for the kinase and nuclear receptor subsets (Table 5) also showed this trend.

The results for Surflex Ringflex are also listed in Table 3. A comparison of the mean RMSDs for Surflex and Surflex Ringflex using the X-ray ligand conformations yielded an interesting result. With ring flexibility turned on, the mean RMSDs were slightly higher, likely due to the fact that the ligand geometry used as input already had the optimal ring conformation encoded and any perturbation would tend to

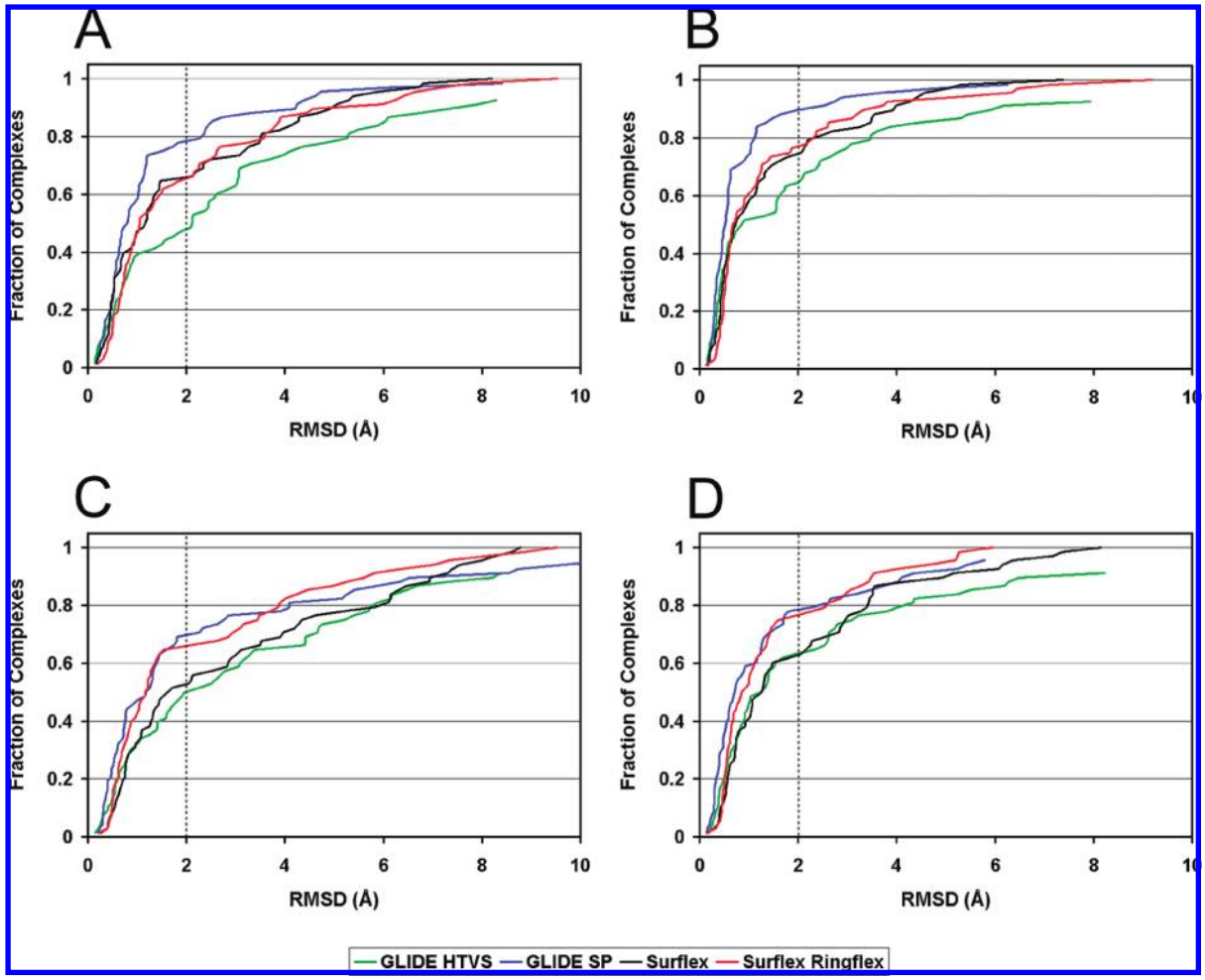

Figure 2. Cumulative distribution plots for cognate ligand docking of 68 protein−ligand complexes using GLIDE and Surflex with parameter tuning: (A) top scoring pose when the X-ray ligand conformation was used as input, (B) best pose when the X-ray ligand conformation was used as input, (C) top scoring pose when the CORINA ligand conformation was used as input, and (D) best pose when the CORINA ligand conformation was used as input. Dotted lines indicate a 2.0 Å RMSD cutoff.

increase the RMSD. This effect was not observed when the CORINA conformations were docked, and Surflex Ringflex yielded lower average RMSDs than Surflex. The difference in mean RMSDs reached statistical significance for the CORINA conformations but not for the X-ray conformations (Table 4). For the kinase subset (Table 5), Surflex Ringflex had higher average RMSDs for the top scoring poses mainly due to a large failure with a single complex (2CL1) relative to Surflex, but the best pose results did show a noticeably smaller mean RMSD. This illustrates, once again, the deficiencies of using mean RMSDs as an evaluation metric. Surflex Ringflex had average RMSDs for the nuclear receptor subset that were consistently much lower than Surflex, but statistical significance was not reached.

Figure 2 shows the fraction of complexes over a distribution of RMSD values for the GLIDE and Surflex alternative parameter settings. When the X-ray ligand conformations were docked, there was a large difference in the number of complexes with poses below the 2.0 Å threshold for GLIDE, with GLIDE SP giving superior results. Little difference was seen between the Surflex and Surflex Ringflex methods. The profiles for GLIDE HTVS and GLIDE SP were similar to the X-ray results at the 2.0 Å threshold when the CORINA conformations were docked. In this case, the Surflex and Surflex Ringflex methods exhibited a distinct separation in the number of complexes with poses below the 2.0 Å threshold, with Surflex Ringflex closely following the GLIDE SP distribution.

Virtual Screening Accuracy. For each of the 40 targets in the DUD data set, the known actives and the target-specific decoys (DUD-self) were docked using each docking program. In theory, this should be a more difficult set of decoys than the entire DUD decoy set, since the physical properties of the DUD-self-decoys are matched to the actives for the target. The curated DUD compound set (actives and decoys) included multiple tautomeric and protonation states for the actives and decoys. Although each of these tautomers was docked, only the highest scoring example of a given compound was retained for analysis. Compounds were then ranked according to their docking scores.

When examining the raw docking results, it was clear that not every compound in the data set was successfully docked by every docking program. This issue of docking failures presents a choice in retrospective virtual screening data analysis; either place all docking failures at the bottom of the ranked list of compounds or remove them from the analysis, since there are no data for these compounds. In this study, compounds that failed to dock were placed at the bottom of the ranked list, with actives and decoys distributed evenly to approximate random selection. The 40 protein targets comprising the DUD, broken down by target, along with the number of successfully docked active and decoy

compounds for each docking program studied, are listed in the Supporting Information. Some of the docking programs, including DOCK, FlexX, and Surflex, consistently returned results for at least 80% of the actives and decoys that were docked. GLIDE HTVS did not reach this docking success threshold for 14 sets of actives and 17 sets of decoys. In most cases, these docking failures occurred for both the actives and decoys for a single target. The docking failures for GLIDE HTVS were concentrated in the nuclear hormone receptor part of the data set, with only one set of actives (for $ER_{agonist}$) reaching an 80% docking success rate. In contrast, GLIDE SP equaled or exceeded the docking success rate compared to GLIDE HTVS for every DUD target and reached an 80% docking success rate more often (failures for 7 sets of actives, 9 sets of decoys). Since the extent of conformational sampling is the only published difference between these methods, it follows that the improved docking success rate for GLIDE SP was likely due to the additional sampling. The results for ICM were slightly better than GLIDE HTVS, with 15 sets of actives and 10 sets of decoys failing to reach 80% threshold, but these failures appeared to be spread more evenly across protein families. The exception to this was the serine proteases, where no target reached the 80% docking success rate. ICM also returned no docking results for any GART actives, which was the only docking program/target combination that failed to dock at least one active compound. It was notable that there were relatively few (21) GART actives compared to most of the other DUD targets and this set also had the lowest degree of diversity, with a quick cosine value of 0.17 when using UNITY fingerprints.[75] If ICM simply had difficulty docking the scaffolds represented in the data set, this would explain the observed result. The results for PhDOCK were interesting in that 2 sets of actives and 37 sets of decoys failed to reach the 80% docking success threshold. While this has the effect of significantly skewing the ratios of actives to decoys across the DUD for this docking program, it also suggests that one of the strengths of PhDOCK is the filtering and elimination of decoys that are not capable of making meaningful contacts with the protein target.

Methodologies for Measuring Virtual Screening Performance. A common metric used when comparing virtual screening results is the enrichment factor (EF)[30]

$$EF^{X\%} = \frac{Actives_{sampled}}{N_{sampled}} \frac{N_{total}}{Actives_{total}}$$

where $Actives_{sampled}$ is the number of actives found at X% of the screened database, $N_{sampled}$ is the number of compounds at X% of the database, N_{total} is the number of compounds in the database, and $Actives_{total}$ is the number of actives in the database. However, the enrichment factor suffers from several deficiencies, making it a less than suitable metric for comparing docking programs.[14,76] Due to its functional form, it explicitly depends on the composition of the data set (i.e. the relative number of actives and decoys present). Use of the DUD eliminates much of the variability in the data set composition, since the active to decoy ratio for each target is approximately 1:36, as long as docking failures are included in the analysis. In addition, the enrichment factor lacks robustness for small values of X, since small changes in the ordering of compounds in this portion of the ranked list can lead to dramatic shifts in enrichment.[26] This unstable behavior occurs precisely where there is the most interest in the results of data analysis, at the earliest portion of the screened database. Even though these issues make it difficult to use the enrichment factor as a metric for comparing docking programs, the mean enrichment factors and mean relative enrichments for the DUD data set, computed with the inclusion of docking failures, have been listed in the Supporting Information.

Use of receiver operating characteristic (ROC) curves is widespread in other fields[77] and is gaining acceptance for the evaluation of modeling methods[17,68,78−80] and structure-based virtual screening.[16,71,76] This metric can be used to effectively differentiate between two populations, so it is suited for evaluating virtual screening performance where one needs to discriminate between active and decoy compounds. It also does not suffer from a lack of robustness with respect to the data set composition (the balance of actives and decoys present) that afflicts the enrichment factor.[26] The area under the curve (AUC) for ROC plots is a powerful metric for virtual screening, corresponding to the probability of correctly ranking an active/decoy ligand pair.[81] AUC values range from 0.0 to 1.0, with 0.5 signifying random selection.

Results for the DUD Data Set. Figure 3 shows the ROC curves for all 40 targets in the DUD data set with each of the docking programs studied. The curves are similar in nature to the recovery curves often seen in virtual screening evaluations, but in the case of ROC curves the *x*-axis indicates the false positive rate rather than the ligand rank. This makes the interpretation of results from the ROC curve somewhat different from a recovery curve; however, the leftmost portion of the curve can still be used to evaluate early recovery of active compounds and random performance is a straight line across the diagonal.

As evidenced by the plots, there is a high degree of variability in virtual screening results, both between targets and between docking programs. Some targets have excellent early recovery results for all docking programs (RXRa, NA), while others have excellent early recovery for only specific programs (DOCK and FlexX for FXa; GLIDE HTVS and Surflex for GPB; FlexX, GLIDE HTVS, ICM, Surflex for DHFR). For other targets, none of the docking programs yielded results significantly different from chance (AChE, AmpC) or gave results inferior to random choice (PDGFrb). Most targets in the study yielded a range of AUC values, indicating that specific docking programs were more or less successful for each target.

The mean AUC values for each docking program are shown in Figure 4 and are tabulated, along with the 95% confidence interval values, in Table 7. Table 8 lists p-values for comparisons between the mean AUCs calculated using the paired *t* test as well as Pearson correlation. Mean difference and correlation corrected 95% confidences are available in the Supporting Information. The correlation in this context only gives a measure of whether two docking programs tended to give similar performance for a given target, not whether the same actives were identified. Further investigation into the correlation of recovery of actives will be presented in a future study. The mean AUC values ranged from 0.55 to 0.72, meaning that all the docking programs were able to choose actives over decoys better than random selection across the entire set of targets, on average. GLIDE

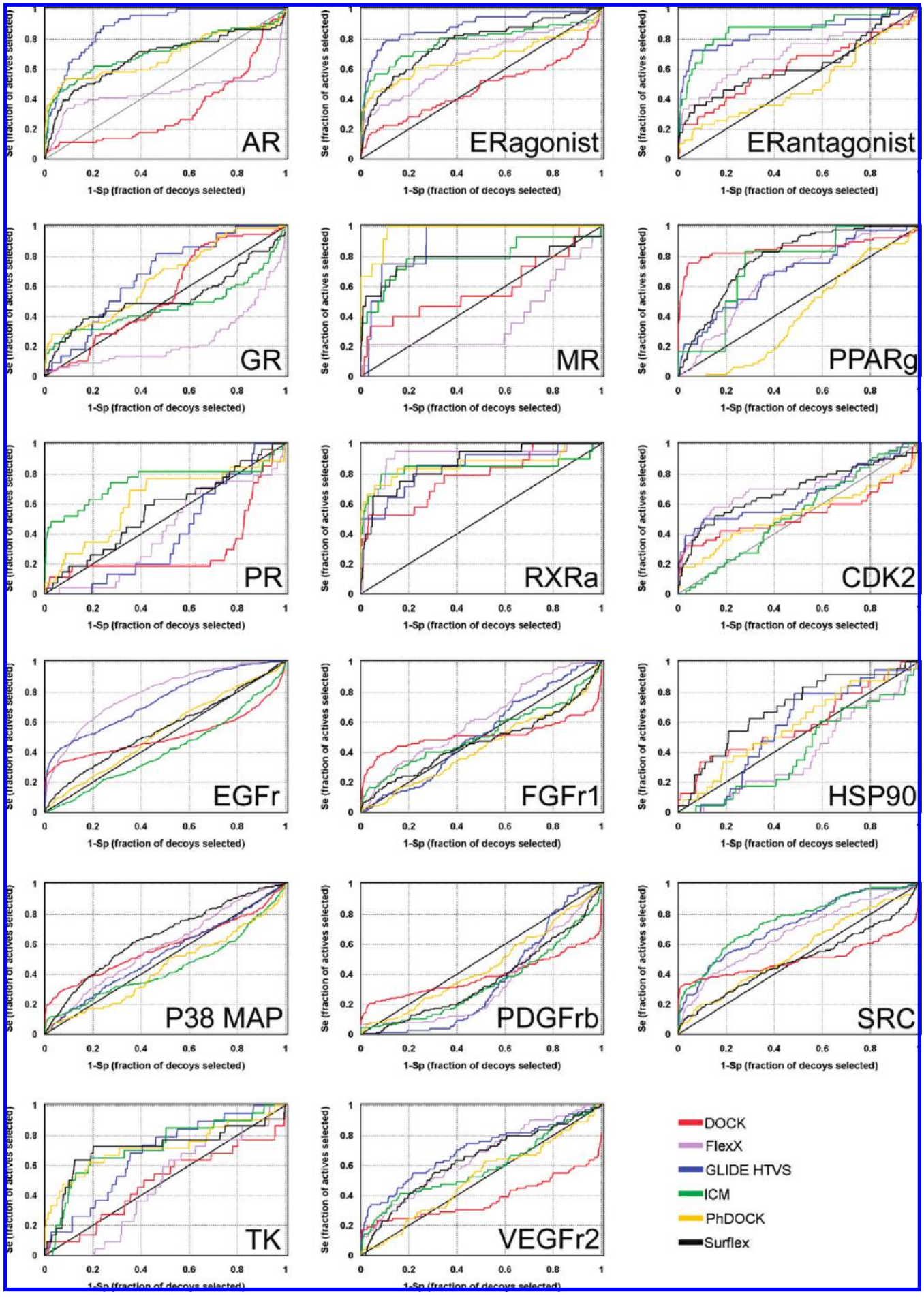

Figure 3. Part 1 of 3. ROC plots for the 40 targets in the DUD data set. Diagonal lines indicate random performance.

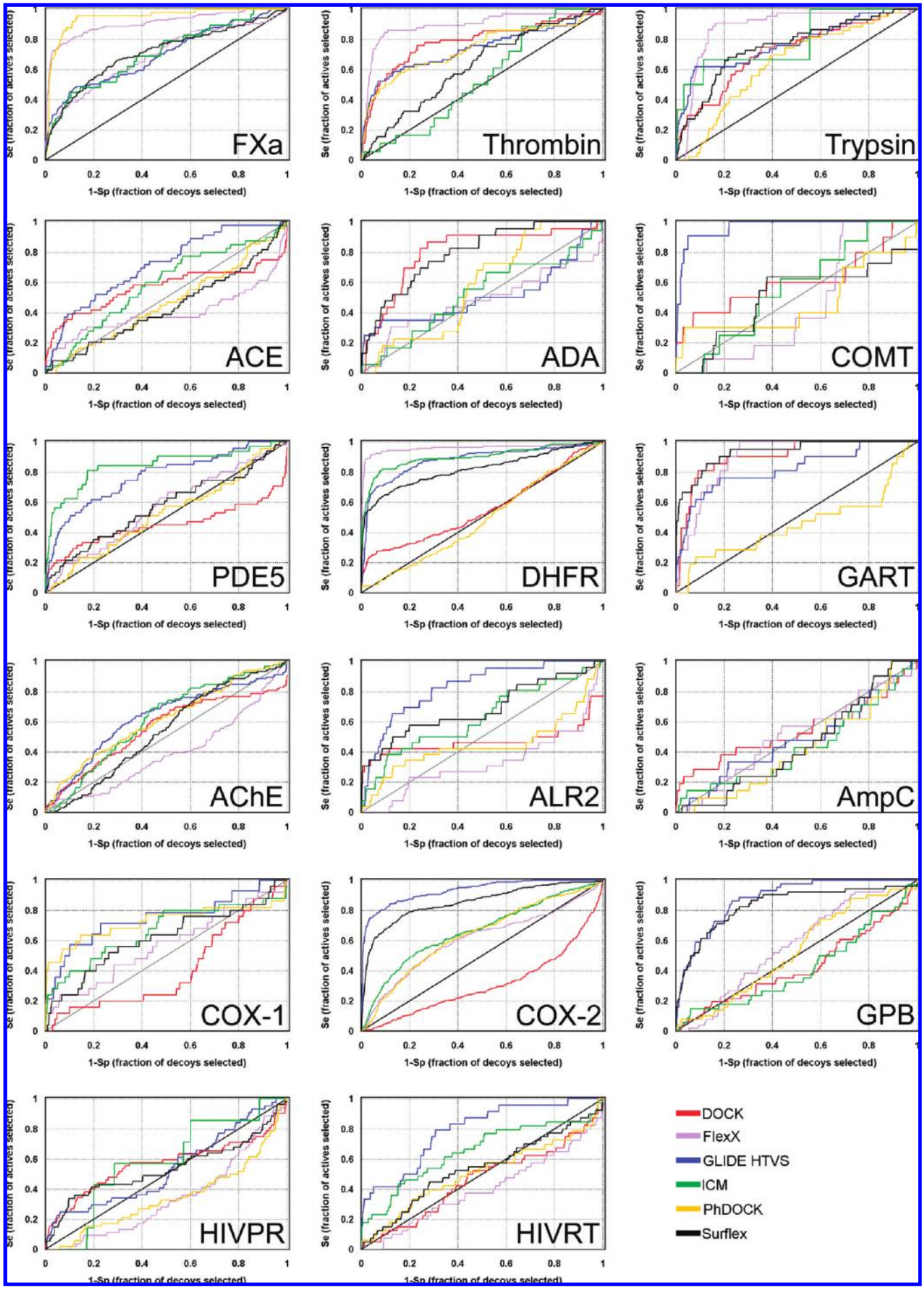

Figure 3. Part 2 of 3.

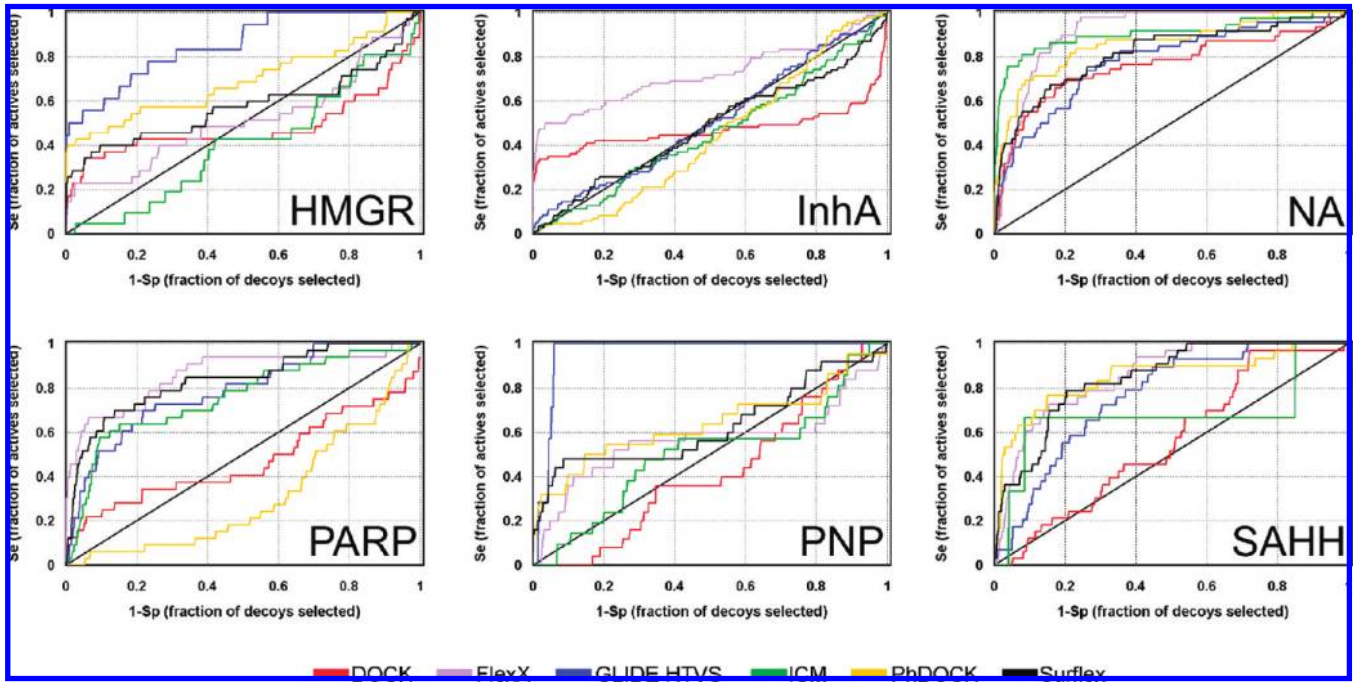

Figure 3. Part 3 of 3.

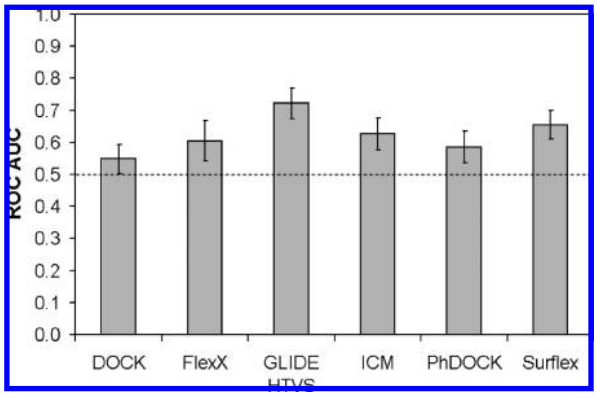

Figure 4. Mean ROC AUC histogram for the DUD data set. Error bars represent 95% confidence intervals. The dotted line at a value of 0.5 indicates random performance.

HTVS yielded the highest mean AUC value at 0.72. This result was significantly better than the AUC values for all other docking programs in the study. Surflex had the second highest mean AUC value at 0.66, which was statistically significant compared to DOCK and PhDOCK. No other mean AUC results were statistically significant with the exception of ICM compared to DOCK. Correlation was generally modest, with Surflex and FlexX having the highest correlation (0.53).

Although high AUC values were often obtained in cases where early recovery of actives was quite good, the reverse was not always true. The FGFr1 and InhA plots in Figure 3 illustrate an interesting feature that occurred often in the DOCK results (for ~25% of the targets) and less often for the other docking programs, involving good early recovery followed by little recovery of actives until much later in the data set. In these cases, the AUC was a poor measure of early recovery of actives and occasionally had an AUC value less than 0.5 for DOCK. A solution for this is to look at metrics that specifically describe early recovery. As previously discussed, the enrichment factor suffers from problems with robustness and dependence on the data set composition. An alternative metric that addresses these deficiencies is the ROC enrichment,[26] which differs from the enrichment factor by referring to the fraction of decoys recovered rather than the fraction of all compounds recovered. While other metrics, such as BEDROC[82] and RIE,[83] have been described in the literature, this analysis was limited to the use of ROC enrichment since it required no adjustable parameter other than choice of false positive rate percentage.

Table 7 contains mean ROC enrichment values across the 40 DUD targets at several points of early recovery. For ROC enrichment values, this corresponds to a low fraction of decoys recovered or low false positive rate. At the 0.5% false positive rate, DOCK gave the best mean ROC enrichment value at 18.8. This notable result can be resolved with the low mean AUC for DOCK if the early and late recovery behavior described in the preceding paragraph is accounted for. As the false positive rate increased, mean ROC enrichment decreased for all docking programs with GLIDE HTVS consistently yielding the highest values and achieving statistical significance against all the other docking programs at the 5.0% and 10.0% false positive rates, with the exception of ICM at the 5.0% level. In fact, the ranking of the docking programs starts to parallel their relative performance based on mean AUC as the false positive rate increases, as can be seen at the 10% level. This indicates that any large variations in the recovery of active compounds that are seen in the very early recovery phase begin to mirror average performance as the false positive rate is increased.

The DUD has significant representation from several protein families of pharmaceutical interest, including nuclear receptors and kinases, and moderate representation for others, including serine proteases, metalloenzymes, and folate enzymes. Some overall trends can be seen for some of these protein families if their mean AUCs and mean ROC enrichments (Table 9) are compared to the entire DUD, even though some of these protein families consist of just a few members. As a whole, the docking programs studied

Table 7. Statistical Results for Virtual Screening Using the DUD, Including Mean ROC AUCs and ROC Enrichments

			mean ROC enrichments[c]				
program	mean ROC AUC[a]	95% CI[b]	0.5%	1.0%	2.0%	5.0%	10.0%
DOCK	0.55	0.50−0.59	18.8	12.3	8.2	4.7	3.0
FlexX	0.61	0.54−0.67	13.7	9.8	7.2	4.4	3.1
GLIDE HTVS	0.72	0.67−0.77	18.9	14.8	10.7	6.5	4.3
ICM[d]	0.63	0.58−0.68	16.9	12.7	8.0	4.6	3.1
PhDOCK	0.59	0.54−0.64	16.9	11.3	7.7	4.1	2.8
Surflex	0.66	0.61−0.70	14.3	11.1	7.9	4.9	3.4
		Tuned Parameters					
GLIDE SP	0.77	0.71−0.82	21.8	16.7	12.2	7.9	5.1
Surflex Ringflex	0.72	0.67−0.77	20.0	16.2	12.0	6.8	4.3

[a] A mean ROC AUC of 0.5 indicates random performance. [b] 95% confidence interval. [c] Mean ROC enrichments were calculated for several early false positive rates. [d] Results for ICM were based on 39 targets, since no actives were returned for GART.

Table 8. Analysis of the Mean ROC AUCs for the DUD Data Set[a]

	DOCK	FlexX	GLIDE HTVS	ICM	PhDOCK	Surflex	GLIDE SP	Surflex Ringflex
DOCK		0.44	0.03	0.19	0.14	0.38	0.18	0.51
FlexX	0.066		0.28	0.42	0.20	0.53	0.29	0.57
GLIDE HTVS	**5.4E-6**	**1.5E-3**		0.42	0.30	0.49	**0.81**	0.43
ICM	**5.7E-3**	0.35	**1.3E-3**		0.33	0.49	0.31	0.45
PhDOCK	0.24	0.61	**3.8E-5**	0.21		0.35	0.11	0.30
Surflex	**1.5E-4**	0.079	**7.3E-3**	0.38	**0.016**		0.51	**0.80**
GLIDE SP	**3.3E-8**	**4.1E-5**	**8.1E-3**	**7.0E-5**	**5.3E-6**	**4.7E-5**		0.59
Surflex Ringflex	**1.0E-8**	**1.5E-4**	0.88	**2.1E-3**	**5.7E-5**	**1.2E-4**	**0.043**	

[a] The upper triangle contains the Pearson correlation, and the lower triangle contains the p-values (calculated using the paired t-test). p-Values in bold are statistically significant (95%). High Pearson correlations (>0.80) are bold. Examples of modified parameter settings are shaded.

produced poorer results for the kinases, with all mean AUCs lower than for the entire DUD. This result is similar to that seen in cognate ligand docking, where kinases tended to have higher RMSD poses compared to the entire data set of protein−ligand complexes. This trend was reversed for the serine proteases and folate enzymes, with all or most of the docking programs yielding higher mean AUCs compared to the entire DUD. Results were mixed for the other protein families. For the nuclear receptors, ICM and PhDOCK tended toward improved AUCs compared to the DUD, while FlexX had a lower mean AUC. DOCK was the only program to have a higher mean AUC for the metalloenzymes, while FlexX, PhDOCK, and Surflex all had lower values than for the DUD. It must also be noted that even though the mean AUCs all tended to be above 0.5, many of the individual AUCs for specific targets were well below 0.5, illustrating the relatively high degree of variability between targets.

Within each protein family, relative performance of each docking program varied compared to the entire DUD. GLIDE HTVS had the highest mean AUC values for the nuclear receptors, kinases, and metalloenzymes, mirroring its ranking with the whole DUD. However, FlexX yielded the highest mean AUC for the serine proteases, and this result was statistically significant compared to DOCK, GLIDE HTVS, and Surflex. For the nuclear receptors, GLIDE HTVS, ICM, and Surflex all gave statistically higher mean AUCs than DOCK and FlexX. GLIDE HTVS and Surflex yielded significantly better mean AUCs than DOCK for the kinase family. Even though there was a wide spread in mean AUC values for the metalloenzymes, none of the results reached statistical significance.

This variability was more evident when inspecting the ROC enrichments. For example, at the 0.5% false positive rate only DOCK and GLIDE HTVS had mean ROC enrichment rates for each protein family that remained close to their averages over the DUD. FlexX had much higher mean ROC enrichments for serine proteases and folate enzymes while performing poorly for metalloenzymes. ICM's performance was improved for nuclear receptors and folate enzymes but suffered for serine proteases. PhDOCK had a higher mean ROC enrichment for the nuclear receptors and lower values for many of the other protein families. Although Surflex's performance suffered for the kinases and serine proteases, it improved greatly for the folate enzymes. It is important to keep in mind the small sample of targets in these subsets and treat these results simply as observed trends within a data set rather than general rules about these docking programs with specific protein families. As the false positive rate is increased to 10.0%, much of the variability observed between protein families at the 0.5% level is

Table 9. Statistical Results for Virtual Screening Using the DUD, Broken down by Protein Family[a]

	DOCK	FlexX	GLIDE HTVS	ICM	PhDOCK	Surflex	GLIDE SP	Surflex Ringflex
ROC AUC								
Entire DUD	0.55	0.61	0.72	0.63	0.59	0.66	0.77	0.72
NHRs	0.56	0.55	0.76	0.74	0.67	0.70	0.81	0.73
Kinases	0.50	0.58	0.61	0.52	0.53	0.58	0.64	0.58
Serine Proteases	0.73	0.89	0.74	0.68	0.78	0.70	0.84	0.89
Metalloenzymes	0.61	0.47	0.74	0.63	0.51	0.58	0.81	0.71
Folate Enzymes	0.73	0.93	0.85	0.90	0.47	0.88	0.97	0.93
ROC Enrichment @ 0.5% FPR								
Entire DUD	18.8	13.7	18.9	16.9	16.9	14.3	21.8	20.0
NHRs	21.7	7.8	13.4	41.5	44.2	15.1	11.7	21.7
Kinases	27.1	10.8	15.0	7.2	7.6	5.5	16.7	10.3
Serine Proteases	17.1	22.2	22.3	4.5	5.7	4.6	31.3	12.1
Metalloenzymes	17.8	2.9	19.5	11.4	3.0	6.2	20.1	12.5
Folate Enzymes	11.4	63.6	28.1	94.5	4.3	77.4	70.9	59.9
ROC Enrichment @ 1.0% FPR								
Entire DUD	12.3	9.8	14.8	12.7	11.3	11.1	16.7	16.2
NHRs	15.9	8.5	14.2	28.8	26.1	14.1	14.2	17.9
Kinases	16.8	6.7	9.8	6.3	6.4	4.7	12.4	7.8
Serine Proteases	12.4	18.6	19.9	14.0	7.8	5.7	25.7	13.1
Metalloenzymes	10.7	2.4	17.5	7.2	2.0	4.4	13.3	14.6
Folate Enzymes	9.1	41.9	20.5	51.0	2.2	43.6	47.1	37.1
ROC Enrichment @ 2.0% FPR								
Entire DUD	8.2	7.2	10.7	8.0	7.7	7.9	12.2	12.0
NHRs	9.6	7.8	11.9	17.4	14.7	9.6	13.9	13.7
Kinases	10.1	4.6	6.1	3.7	4.0	3.2	8.7	5.4
Serine Proteases	9.0	15.0	13.2	9.0	14.6	5.3	17.5	15.1

Table 9. Continued

	DOCK	FlexX	GLIDE HTVS	ICM	PhDOCK	Surflex	GLIDE SP	Surflex Ringflex
ROC Enrichment @ 2.0% FPR								
Metalloenzymes	8.5	1.2	13.4	5.9	1.2	3.0	11.3	8.6
Folate Enzymes	10.1	29.0	15.6	29.4	1.1	27.7	30.4	31.7
ROC Enrichment @ 5.0% FPR								
Entire DUD	4.7	4.4	6.5	4.6	4.1	4.9	7.9	6.8
NHRs	5.6	4.4	7.4	8.5	6.9	6.2	8.6	7.8
Kinases	4.7	2.8	3.7	2.5	2.3	2.5	5.2	3.0
Serine Proteases	6.5	11.2	7.5	5.4	7.5	4.3	11.4	10.7
Metalloenzymes	5.3	1.5	8.0	3.3	1.9	2.4	7.5	4.1
Folate Enzymes	7.7	12.0	11.4	14.8	0.5	12.3	16.1	16.3
ROC Enrichment @ 10.0% FPR								
Entire DUD	3.0	3.1	4.3	3.1	2.8	3.4	5.1	4.3
NHRs	3.1	2.9	4.7	5.0	4.3	4.0	5.5	4.6
Kinases	2.9	2.1	2.4	1.9	1.7	2.5	3.4	2.4
Serine Proteases	4.1	7.7	5.3	3.4	4.8	2.8	7.7	7.5
Metalloenzymes	3.5	1.4	5.3	2.1	1.7	1.9	5.4	2.9
Folate Enzymes	5.4	7.3	6.7	8.2	1.6	7.0	9.4	8.6

[a] Examples of modified parameter settings are shaded.

reduced, and the results for each docking program appear much smoother relative to the entire DUD.

As with the cognate ligand docking portion of this study, the effects of simple parameter modifications on the virtual screening results were also examined. The same methods were compared, and the GLIDE SP and Surflex Ringflex results are appended to Tables 7, 8, and 9. As shown in Figure 5, the mean AUC was improved for GLIDE SP compared to GLIDE HTVS and for Surflex Ringflex compared to Surflex, and both of these results were statistically significant. This trend in the mean AUCs was consistent even with the DUD broken down by protein family, with Surflex Ringflex giving significantly higher mean AUCs than Surflex for the nuclear receptors, serine proteases, and metalloenzymes. The mean ROC enrichment rates showed the same type of improvement as well, though the difference was more dramatic for Surflex Ringflex than for GLIDE SP. At all false positive rates other than 0.5%, the mean ROC enrichments for Surflex Ringflex were significantly higher than Surflex. Surflex Ringflex was also capable of outperforming GLIDE HTVS at all early mean ROC enrichment levels other than 10.0% (where these methods were equivalent), which contrasts with the Surflex results that were consistently below those of GLIDE HTVS. There was more variability when the mean ROC enrichments were broken down by protein family, but the same trend tended to hold.

CONCLUSIONS

This study showed that docking programs, in general, can successfully generate poses that closely resemble the X-ray

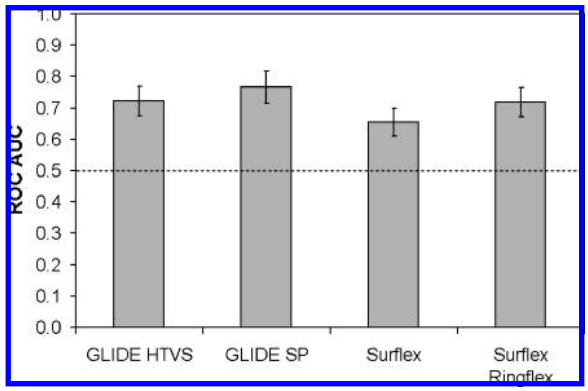

Figure 5. Mean ROC AUC histogram for the DUD data set with parameter tuning of GLIDE and Surflex. Error bars represent 95% confidence intervals. The dotted line at a value of 0.5 indicates random performance.

conformation and are also able to identify active compounds in a pool of decoys. These results hold even when no software parameters are modified, with the docking programs being applied "out of the box". Attempts were made to limit the bias in this study by using a standalone program (CORINA) to generate 3D ligand conformations, keeping protein heavy atom coordinates fixed, and minimizing the hydrogen atoms on the protein without a ligand present in the binding site. However, most of the protein preparation was performed using the Maestro interface, which is the preferred method for GLIDE, and while care was taken to minimize potential bias one cannot be certain that this did not affect the docking and virtual screening results in a subtle way.

While certain docking programs, such as GLIDE and Surflex, were capable of superior performance in both cognate ligand docking and virtual screening, other tools appeared to favor specialized applications. For instance, ICM was quite successful in generating and identifying low RMSD poses in cognate ligand docking but had middling performance in virtual screening. On the other hand, Surflex's virtual screening results (in default mode) were better than many of the other docking programs, but its success rate for cognate ligand docking was essentially average without ring flexibility enabled.

There were also trends identified in docking and virtual screening success rates within specific protein families. For kinases, GLIDE (SP and XP for cognate ligand docking and HTVS for virtual screening), ICM, and Surflex performed best overall. Although the differences between the cognate docking results for the docking programs were not statistically significant for the nuclear receptors, GLIDE (SP and XP) and ICM had the lowest mean RMSDs. However, in virtual screening, GLIDE HTVS, Surflex, and ICM all outperformed DOCK and FlexX for this protein family. FlexX had the highest mean AUC for the serine proteases, a result that was significant compared to DOCK, GLIDE HTVS, and Surflex. Differences in performance for these various protein families could be due to the composition of the training sets used to develop docking programs, different intended goals for the development of these tools (pose prediction vs database screening), or scoring functions that have a bias for particular physical properties of both protein binding sites and ligands.

Significant effects on docking and virtual screening accuracy were also seen when software parameters were modified. With GLIDE, the SP method performed better than the HTVS method in both of these situations. Surflex Ringflex also gave results that were superior to Surflex in default mode. These results were consistent with the notion that GLIDE SP and Surflex Ringflex are more accurate versions of the GLIDE HTVS and Surflex (default) methods. Even though an extremely limited set of options was explored with two of the docking programs, it is expected that similar results could be obtained for any docking program with a robust set of adjustable parameters.

The docking programs evaluated in this study exhibited an array of strengths, depending on the application. Since the majority of the docking and virtual screening results presented were obtained using default parameters, these conclusions should be taken as more of a rough guide when selecting a tool for a specific situation rather than a firm statement of expected performance. Updated versions of some docking programs became available during the course of this evaluation, and the results of any improvements are therefore not reflected in this study. Expert knowledge is also critical when optimizing software parameters for each new protein target or chemical series encountered. Before engaging on a major docking or virtual screening exercise, some validation of a particular docking program/target combination is certainly warranted where X-ray structures and known ligands exist. Even though statistically significant differences between the docking programs were seen it must be noted that these results are still dependent on the data sets and targets used in the study, although for virtual screening the DUD is quite a large, generalized data set. Deviations from this average behavior is to be expected when these programs are applied to other targets, as seen in the protein family analyses, or with parameter changes intended to alter the balance between computational speed and accuracy.

ACKNOWLEDGMENT

The authors thank Tianhui Zhou for assistance with statistical data analysis. We also thank Diane Joseph-McCarthy, Natasja Brooijmans, and Rayomand Unwalla for discussions on experimental design and docking methodology as well as the reviewers for their thoughtful suggestions and insights. The authors also acknowledge BioSolveIT, MolSoft, and Tripos for generously providing demo versions of their software.

Supporting Information Available: Tables of mean differences and correlation corrected confidence intervals for cognate ligand docking and virtual screening, docking success rates for the DUD data set, mean enrichment factors, and mean relative enrichment factors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301−D303.

(2) Lang, P. T.; Aynechi, T.; Moustakas, D.; Shoichet, B.; Kuntz, I. D.; Brooijmans, N.; Oshiro, C. M., Molecular docking and structure-based design. In *Drug Discovery Research: New Frontiers in the Post-*

Genomic Era; Huang, Z., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2007; pp 3−23.
(3) Muegge, I.; Oloff, S. Advances in virtual screening. *Drug Discovery Today: Technol.* **2006**, *3*, 405–411.
(4) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(5) 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.
(6) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.
(7) Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N. Evaluation of docking programs for predicting binding of Golgi a-mannosidase II inhibitors: a comparison with crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 160–176.
(8) 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.
(9) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(10) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47–57.
(11) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962–976.
(12) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11–22.
(13) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.
(14) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(15) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.
(16) Deng, W.; Verlinde, C. L. M. J. Evaluation of Different Virtual Screening Programs for Docking in a Charged Binding Pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2010–2020.
(17) Kellenberger, E.; Foata, N.; Rognan, D. Ranking Targets in Structure-Based Virtual Screening of Three-Dimensional Protein Libraries: Methods and Problems. *J. Chem. Inf. Model.* **2008**, *48*, 1014–1025.
(18) Sheridan, R. P.; McGaughey, G. B.; Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 257–265.
(19) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2004**, *17*, 755–763.
(20) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225–242.
(21) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235–249.
(22) 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.
(23) Onodera, K.; Satou, K.; Hirota, H. Evaluations of Molecular Docking Programs for Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1609–1618.
(24) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325–332.
(25) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
(26) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(27) Liebeschuetz, J. W. Evaluating docking programs: keeping the playing field level. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 229–238.
(28) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.
(29) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection-What can we learn from earlier mistakes. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.
(30) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
(31) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(32) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.
(33) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 191–192.
(34) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457–471.
(35) 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.
(36) *CORINA, Version 1.82*; Molecular Networks GmbH: Erlangen, Germany, 1997.
(37) *Maestro, Version 7.5.116*; Schrodinger, LLC: Portland, OR, U.S.A., 2006.
(38) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13*, 380–397.
(39) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175–1189.
(40) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601–619.
(41) Richards, F. M. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
(42) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722–729.
(43) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(44) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 228–241.
(45) Boehm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
(46) Boehm, H. J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593–606.
(47) Klebe, G. The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. *J. Mol. Biol.* **1994**, *237*, 212–235.
(48) 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.
(49) 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.
(50) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
(51) 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. Des.* **1997**, *11*, 425–445.
(52) 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.
(53) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.

(54) Totrov, M.; Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Bioinf.* **1998**, (Suppl. 1), 215–220.

(55) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17*, 616–641.

(56) Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* **1992**, *96*, 6472–6484.

(57) Abagyan, R.; Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235*, 983–1002.

(58) Joseph-McCarthy, D.; Thomas, B. E. I. V.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 172–188.

(59) Joseph-McCarthy, D.; McFadyen, I. J.; Zou, J.; Walker, G.; Alvarez, J. C. Pharmacophore-based molecular docking: A practical guide. *Drug Discovery Ser.* **2005**, *1*, 327–347.

(60) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4. 0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428.

(61) Rush, T. S., III.; Manas, E. S.; Tawa, G. J.; Alvarez, J. C. Solvation-based scoring for high-throughput docking. *Drug Discovery Ser.* **2005**, *1*, 249–277.

(62) Joseph-McCarthy, D.; Alvarez, J. C. Automated generation of MCSS-derived pharmacophoric DOCK site points for searching multiconformation databases. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 189–202.

(63) *OMEGA, Version 2.2.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2007.

(64) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **1999**, *20*, 720–729.

(65) *Szybki, Version 1.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2004.

(66) *ZAP, Version 1.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2006.

(67) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(68) Thompson, D. C.; Humblet, C.; Joseph-McCarthy, D. Investigation of MM-PBSA rescoring of docking poses. *J. Chem. Inf. Model.* **2008**, *48*, 1081–1091.

(69) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(70) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281–306.

(71) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(72) Jain, A. N. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.

(73) *OEChem, Version 1.4.0*; OpenEye Scientific Software, Inc.: Sante Fe, NM, U.S.A., 2006.

(74) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.

(75) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.

(76) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(77) Swets, J. A.; Dawes, R. M.; Monahan, J. Better DECISIONS through SCIENCE. *Sci. Am.* **2000**, *283*, 82.

(78) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199–213.

(79) Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Rault, S. The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/Biodegradability Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1043–1052.

(80) Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47*, 947–961.

(81) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.

(82) 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.

(83) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395–1406.

CI900056C