Benchmarking Sets for Molecular Docking

Niu Huang, Brian K. Shoichet,* and John J. Irwin*

Department of Pharmaceutical Chemistry, University of California San Francisco, QB3 Building, 1700 4th Street, Box 2550, San Francisco, California 94143-2550

Received July 16, 2006

Ligand enrichment among top-ranking hits is a key metric of molecular docking. To avoid bias, decoys should resemble ligands physically, so that enrichment is not simply a separation of gross features, yet be chemically distinct from them, so that they are unlikely to be binders. We have assembled a directory of useful decoys (DUD), with 2950 ligands for 40 different targets. Every ligand has 36 decoy molecules that are physically similar but topologically distinct, leading to a database of 98 266 compounds. For most targets, enrichment was at least half a log better with uncorrected databases such as the MDDR than with DUD, evidence of bias in the former. These calculations also allowed 40 × 40 cross-docking, where the enrichments of each ligand set could be compared for all 40 targets, enabling a specificity metric for the docking screens. DUD is freely available online as a benchmarking set for docking at http://blaster.docking.org/dud/.

Introduction

Although molecular docking screens of chemical databases are widely used for ligand discovery,[1−7] the method retains important weaknesses.[8−13] A testament to these is the criterion by which docking screens are evaluated: the enrichment of annotated ligands from among a large database of presumed nonbinding "decoy" molecules. In these retrospective calculations, the enrichment factor is the concentration of the annotated ligands among the top-scoring docking hits compared to their concentration throughout the entire database. Other possible metrics, such as the magnitude of the docking energies, or even monotonic rank order among the ligands, are only used occasionally and in restricted sets; in the general case, they remain unreliable because of the many approximations used in docking. Thus, the success of a docking screen is evaluated by its capacity to enrich the small number of known active compounds in the top ranks of a screen from among a much greater number of decoy molecules in the database.[14−23]

The relationship of the decoy molecules to the ligands is critical in assessing enrichment factors in docking screens. Docking scoring functions can depend on molecular size.[24−26] For instance, Verdonk and colleagues[27] have observed that if there are significant differences in size distribution between ligands and decoys, docking enrichments can appear to be artificially good, and the same is undoubtedly true for other physical features. The database decoys should thus resemble the physical properties of the annotated ligands well enough so that enrichment is not simply a separation of trivial physical features. The decoys nevertheless should be chemically distinct from the ligands so that they are likely to be, in fact, nonbinders.

Investigators have assembled sets of ligands and presumed decoys for numerous targets and used them to evaluate docking performance based on enrichment. Rognan and colleagues made an important contribution toward this end with the introduction of a set of 990 randomly chosen molecules combined with 10 thymidine kinase (TK)[a] and 10 estrogen receptor (ER) antagonists, which were subsequently used in several studies.[15,19,20,23,28−30] More recently, Jain and colleagues introduced a set of 1000 random druglike compounds to complement the Rognan set and combined those with 252 ligands from 27 protein targets to evaluate docking enrichment factors.[23] Several other groups, including ourselves, have used the 100 000 molecule MDL Drug Data Report database (MDDR, Elsevier MDL, San Leandro CA) as a source of both ligands and decoys.[31−34] Each of these approaches has drawbacks. None of these sets of molecules has been adjusted so that the physical properties of the ligands are matched by those of the decoys. Indeed, in both the Rognan set and its derivatives the annotated ligands and the presumed decoys differ greatly in their physical properties, making enrichment factors calculated with these sets open to bias (see Results). Whereas there is less room for bias in the MDDR database, and statistical variance is less likely here than in the smaller decoy sets, differences between ligand and decoy sets lead to significant enrichment-factor bias (see Results). The MDDR has the further disadvantage of being a nonpublic access database, which was an advantage of the Rognan and Jain sets.

We were interested in developing large benchmarking sets to evaluate docking screening calculations. We wanted these sets to cover a large number of proteins so as to offer a reliable view of how docking might perform on typical and interesting targets. We also wanted these sets to be publicly available, so that docking programs could be compared broadly in "apples

* To whom correspondence should be addressed. B.K.S: phone, 415-514-4126; fax, 415-514-4260; E-mail, shoichet@cgl.ucsf.edu. J.J.I: phone, 415-514-4127; fax, 415-514-4260; e-mail, jji@cgl.ucsf.edu.

[a] Abbreviations: DUD, directory of useful decoys; EF, enrichment factor; MDDR, MDL Drug Data Report; Tc, Tanimoto coefficient; ROC, receiver operating characteristic; ACE, angiotensin-converting enzyme; AChE, acetylcholinesterase; ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR, androgen receptor; CDK2, cyclin-dependent kinase 2; COMT, catechol *O*-methyltransferase; COX-1, cyclooxygenase-1; COX-2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFr, epidermal growth factor receptor; ER, estrogen receptor; FGFr1, fibroblast growth factor receptor kinase; FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPB, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethylglutaryl-CoA reductase; HSP90, human heat shock protein 90; InhA, enoyl ACP reductase; MR, mineralocorticoid receptor; NA, neuraminidase; P38 MAP, P38 mitogen activated protein; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase 5; PDGFrb, platelet derived growth factor receptor kinase; PNP, purine nucleoside phosphorylase; PPARg, peroxisome proliferator activated receptor γ; PR, progesterone receptor; RXRa, retinoic X receptor α; SAHH, S-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase SRC; TK, thymidine kinase; VEGFr2, vascular endothelial growth factor receptor; ATP, adenosine-5′-triphosphate; β-GAR, β-glycinamide ribonucleotide; NAD(P)-(H), nicotinamide adenine dinucleotide (phosphate)-(reduced); PLP, pyridoxal-5′-phosphate.

10.1021/jm0608356 CCC: $33.50
Published on Web 10/26/2006

to apples" comparisons. We wanted the database to be large enough to have decoys and ligands physically matched so as to be as free from physical and statistical features as possible. We began with 2950 ligands for 40 different target proteins taken from the literature; each set of ligands for each protein had tens to hundreds of molecules in it. For each ligand in each set, 36 molecules were chosen from the "druglike" subset of the ZINC database of commercially available compounds.[35] Each of these 36 resembled the particular ligand in physical properties, such as molecular weight, cLogP, and number of hydrogen bonding groups, but differed from the ligand topologically. This resulted in a database of 95 316 decoy molecules whose physical properties closely matched those of the 2950 ligands that they were chosen to counterpoint. This "directory of useful decoys" (DUD) was docked against the 40 protein targets, using an automated docking engine that required little or no user guidance.

Here we report on the enrichments resulting from this large, bias-corrected database and compare these to those from both small, uncorrected decoy sets[15,23] and to the large MDDR database. Our results suggest that DUD provides a more stringent test with which to evaluate virtual screening performance. Even using the MDDR as decoys has enrichment factors half a log better than those using DUD with exactly the same docking procedure, speaking to enrichment factor biases even in large, uncorrected databases. These calculations also allowed for a 40 × 40 cross-docking, where the enrichments of each ligand set were compared for all forty targets using the same background decoys. These cross-docking results suggest a specificity metric to evaluate docking screens. The usefulness of these targets, ligands, and decoys as community benchmarking sets for docking will be considered here, as will the prospects for the automated docking pipeline that was deployed in these studies. The benchmarking sets, including protein structures, docking energy grids and input files, and the full DUD database are available at blaster.docking.org/dud/.

Method

Protein Target Selection and Ligand Collection. Forty protein targets were selected on the basis of the availability of annotated ligands, crystal structures, and, often, previous docking studies. We used the structure used in previous docking studies if one was available, and otherwise we used the best complex available as judged by resolution and the absence of errors. All of these proteins have ligand-bound X-ray crystal structures available in the Protein Data Bank (PDB),[36] with the exception of the PDGFrb and VEGFr2 kinases. We organized these targets into six classes: nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes (Table 1). The number of ligands varies from 12 (~ 0.01% of database) to 416 (~ 0.4% of database). A total of 2950 ligands were included overall.

Most of these targets have been studied previously by experimental methods and computational approaches. Among them, estrogen receptor (ER)[15,20,23,28,29,37−39] and thymidine kinase (TK)[15,19,20,23,28,30,40] have been extensively used to benchmark and evaluate different docking methods and scoring functions. Enrichment studies were also published on several of the other systems, including CDK2,[20,41] P38 MAP kinase,[18,20,32,39] thrombin,[20,31,39,41] factor Xa,[38,42] HIV protease,[18,20,40] DHFR,[31,40] neuraminidase,[39] aldose reductase,[31,43] HIV-RT,[20,30] AChE,[31,38] and COX-2.[20,39,44]

DUD Generation. DUD was created as follows (Figure 1). The 2950 annotated ligands were seeded among 3.5 million Lipinski-compliant molecules from the ZINC database of commercially available compounds (version 6, December 2005).[35] Chiral annotated ligands were prepared in the correct stereochemical form if known. Feature key fingerprints were calculated using the default type 2 substructure keys of CACTVS,[45] and the fingerprint-based similarity analysis was performed with the program SUBSET.[46] Substructure keys are bit strings where 1 represents the presence of a particular functional group. Compounds with Tanimoto coefficient (Tc) less than 0.9 to any annotated ligand were selected, excluding chirality duplicates (we note that a Tc less than 0.9 for CACTVS type 2 fingerprints roughly corresponds to a Tc less than about 0.7 for the widely used Daylight fingerprints; see Results). This reduced the ZINC compounds to 1.5 million molecules topologically *dissimilar* to the ligands. The program QikProp (Schrodinger, LLC, New York, NY) was used to calculate 32 physical properties of all the annotated ligands and selected ZINC compounds from the previous step, and QikSim (Schrodinger, LLC, New York, NY) was applied to prioritize ZINC compounds

Table 1. Enrichments of the Annotated Ligands Using the Decoys in DUD for Forty Targets by Docking[a]

protein	PDB code	resolution (Å)	no. of ligands[b]	no. of decoys	EF_{max}	EF_1	EF_{20}
Nuclear Hormone Receptors							
AR	1xq2	1.9	74 (a,b)	2630	60.2	33.5	3.8
$ER_{agonist}$	1l2i	1.9	67 (a−c)	2361	29.6	19.2	4.5
$ER_{antagonist}$	**3ert**	**1.9**	**39 (a−d)**	**1399**	**101.6**	**12.7**	**1.3**
GR	1m2z	2.5	78 (a)	2804	31.7	8.9	1.4
MR	2aa2	1.9	15 (a)	535	330.0	46.2	3.7
PPARg	1fm9	2.1	81 (a)	2910	1.0	0.0	0.0
PR	1sr7	1.9	27 (a)	967	2.9	0.0	2.0
RXRa	1mvc	1.9	20 (a)	708	148.5	24.8	2.2
Kinases							
CDK2	1ckp	2.1	50 (e,f)	1780	19.8	13.9	1.4
EGFr	1m17	2.6	416 (g)	14914	3.8	2.1	2.4
FGFr1	1agw	2.4	118 (g)	4216	1.0	0.0	0.2
HSP90	1uy6	1.9	24 (h)	861	10.8	8.6	2.0
P38 MAP	**1kv2**	**2.8**	**234 (g)**	**8399**	**4.1**	**2.1**	**2.4**
PDGFrb	model	n/a	157 (g)	5625	1.2	0.0	0.6
SRC	2src	1.5	162 (g)	5801	3.1	1.2	1.5
TK	**1kim**	**2.1**	**22 (a,d,i)**	**785**	**63.0**	**54.0**	**5.0**
VEGFr2	1vr2	2.4	74 (j)	2647	2.2	1.3	1.4
Serine Proteases							
FXa	1f0r	2.7	142 (e,f,k)	5102	34.9	14.6	3.8
thrombin	1ba8	1.8	65 (e,l,m)	2294	18.3	13.7	2.9
trypsin	1bju	1.8	43 (e,l)	1545	22.5	22.5	2.6
Metalloenzymes							
ACE	1o86	2.0	49 (a,m)	1728	141.4	40.4	3.7
ADA	**1stw**	**2.0**	**23 (a,e)**	**822**	**21.5**	**12.9**	**2.4**
COMT	1h1d	2.0	12 (a)	430	11.8	0.0	3.3
PDE5	1xp0	1.8	51 (f)	1810	29.1	11.8	2.3
Folate Enzymes							
DHFR	3dfr	1.7	201 (m)	7150	28.7	21.7	3.5
GART	1c2t	2.1	21 (n)	753	70.7	42.4	3.3
Other Enzymes							
AChE	1eve	2.5	105 (a,e,m)	3732	3.1	1.9	2.0
ALR2	**1ah3**	**2.3**	**26 (o)**	**920**	**76.2**	**38.1**	**2.3**
AmpC	1xgj	2.0	21 (p)	734	23.6	17.1	4.7
COX-1	1p4g	2.1	25 (i)	850	9.9	4.0	1.6
COX-2	1cx2	3.0	349 (c,f,m)	12491	29.1	20.1	3.3
GPB	1a8i	1.8	52 (e,m)	1851	28.6	22.8	4.1
HIVPR	1hpx	2.0	53 (a,e)	1888	9.3	3.7	2.2
HIVRT	1rt1	2.6	40 (q)	1439	49.5	5.0	3.0
HMGR	1hw8	2.1	35 (a,i)	1242	198.0	33.9	2.1
InhA	**1p44**	**2.7**	**85 (r)**	**3043**	**1.0**	**0.0**	**0.3**
NA	1a4g	2.2	49 (c,e,i)	1745	60.6	20.2	3.3
PARP	1efy	2.2	33 (s)	1178	6.3	6.0	3.6
PNP	1b8o	1.5	25 (e,t)	884	158.4	31.7	4.4
SAHH	1a7a	2.8	33 (i)	1159	120.0	78.0	5.0

[a] Six representative targets (in bold) are discussed in more detail in the text. [b] Annotated ligands were collected from (a) KiBank,[67] (b) NCTR data set,[68] (c) Stahl data set,[16] (d) from ref 15, (e) PDBbind database,[69] (f) Jorissen/Gilson data set,[70] (g) Kinase inhibitor data set,[32] (h) refs 71 and 72, (i) PubChem (http://pubchem.ncbi.nlm.nih.gov), (j) refs 73−75, (k) Jacobsson test set,[38] (l) Bohm serine protease inhibitor data set,[76] (m) Sutherland QSAR test set,[77] (n) ref 78, (o) ref 79, (p) refs 5, 80, and 81, (q) ref 82, (r) contributed by Dr. Xin He (UCSF, personal communication), (s) ref 83, and (t) ref 84.

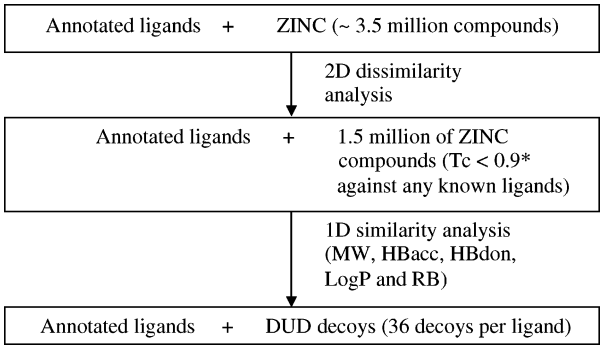


Figure 1. The schematic description of the procedure to generate DUD: Molecular weight (MW), number of hydrogen bond acceptors (HBacc), number of hydrogen bond donors (HBdon), number of rotatable bonds (RB). *We note that a Tc less than 0.9 for CACTVS type 2 fingerprints roughly corresponds to a Tc less than about 0.7 for the Daylight fingerprints.

possessing similar properties to any of the ligands. A weight of 4 was used to emphasize the druglike descriptors (molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rotatable bonds, and log *P*), and a weight of 1 was used for the number of important functional groups (amine, amide, amidine, and carboxylic acid), and the rest of the descriptors were ignored (weight 0) during the similarity analysis procedure. Thirty-six decoy compounds were selected for each ligand, leading to a total of 95 316 decoys that were physically similar but topologically dissimilar to the 2950 annotated ligands. The total number of decoys is less than 36 times the number of annotated ligands because some ligands had the same decoys.

Dockable Database Preparation. Molecules were prepared for docking using the latest version of the ZINC protocol.[35] Briefly, molecules were converted from 2D SDF to isomeric SMILES using OEChem (OpenEye Scientific Software, Santa Fe, NM). An initial 3D structure was generated with Corina (Molecular Networks GmbH). A protonated form of each molecule at pH 7.0 was calculated with LigPrep (Schrodinger, LLC, New York, NY) with additional protonated and tautomeric forms calculated in the range of pH 5.75−8.25 using modified versions of LigPrep's parameter files. For each protonated form, we again used Corina to obtain a 3D model and then used AMSOL to calculate partial atomic charges and atomic desolvation energies.[47] We used Omega (OpenEye Scientific Software, Santa Fe NM) to enumerate accessible conformations; ring conformations calculated by Corina were preserved. AMSOL and Omega results were combined into a single "flexibase" format file using our program Mol2db.[33] All new parameter files used in this process (rules.txt, tautomer_list, ionizer.ini, and torlib.txt) are available in the Supporting Information.

Overall Virtual Screening Strategy. To undertake docking screens against 40 targets, it was important to automate our procedures as much as possible (Supporting Information, Figure S1). Most of the labor-intensive steps formerly performed manually have been automated, including most of the binding site preparation, sphere or "hot spot" generation, scoring grid calculation, docking calculation, and data analysis. For simplicity, our automated procedure removes all water molecules, including structural waters, by default. We describe docking results achieved with or without expert intervention as "semiautomated" or "automated", respectively. For both semiautomated and automated procedures, the only input requirements were a protein structure file and a specification of the protein-binding site. Expert intervention during protein structure preparation or binding site identification significantly improved the docking results for 13 out of the 40 targets (see Results). The fully automated procedure was used for all 40 targets, the semiautomated procedure being attempted only when docking enrichment from the fully automated procedure was poor.

Manual Preparation. For some targets, protein structure preparation involves steps that are challenging to automate, such as differentiating cofactors from ligands, parametrizing those cofactors, perceiving structural water molecules, identifying and parametrizing metal ions involved in ligand binding, correctly assigning the protonation state on binding site residues (e.g. histidines and cysteines), and selecting among disordered residues.

Parametrizing the cofactor is challenging to automate. Cofactors were present for the following targets: TK, SO_4; DHFR, NADPH; GART, β-GAR; ALR2 and InhA, NADP+; GPB, PLP; PNP, PO_4; and SAHH, NAD+. In these cases, we treated the cofactors as part of the target, manually preparing their parameters for the van der Waals (vdW) and electrostatic energy calculations. Once a parameter file has been prepared for a cofactor, the scripts can recognize it on the basis of its PDB residue name, and it becomes part of the automated procedure in future runs.

For control calculations, the crystallographic ligand was also prepared manually for docking using SYBYL.[48] In the case of PDGFrb kinase, where no crystal structure was available and a modeled structure was used,[49,50] the cognate ligand was obtained from the X-ray crystal structure of c-Kit kinase, the homology modeling template. Similarly, only an uncomplexed apo structure was available for VEGFr1 kinase, so its native ligand was obtained by superimposing FGFr1 kinase, a homologue with high sequence and structural identity.

In target systems with large ligands spanning more than one pocket, it is helpful to specify that part of the ligand most intimately involved in binding. Such a fragment is presented to the automated scripts as an individual file that can be recognized as the reference state for generating the docking spheres or "hot spots". Other special measures include manually redistributing the partial atomic charges of polar atoms in critical binding site residues to increase polarity and thus favoring polar ligands, as described previously.[5,34,51]

Automated Steps. The automated docking pipeline begins with the receptor structure file and its cocrystallized ligand or a manually curated specification of the binding site. All tasks, including sphere generation, scoring grid and docking calculations, and analysis of enrichment, are driven automatically (Figure S1). The scheduling system Condor (University of Wisconsin, Madison, WI) was used to manage jobs on our Linux cluster.

Binding site residues are identified as those being within 12 Å of any heavy atom of the crystallographic ligand or the residues used to define the site, using the program FILT (from the DOCK3.5 distribution). The solvent-accessible molecular surface[52] of the protein binding site is then calculated with the program DMS[53] using a probe radius of 1.4 Å. Receptor-derived spheres are calculated using the program SPHGEN (part of the UCSF DOCK suite),[54] while the ligand-derived spheres are simply generated from the positions of the heavy atoms of the crystallographic ligand, if available. If the molecular fragment file is present, the ligand-derived spheres are created from the molecular fragment instead of using the entire ligand structure. The matching spheres, required for orientation of the ligand in the binding site, are obtained by augmenting the ligand-derived spheres with receptor-derived spheres. Spheres furthest away from ligand-derived spheres, furthest from the centroid of the remaining spheres, too close to receptor atoms, or too close to each other are removed iteratively until the total number of spheres is 35 or less. Spheres are labeled for chemical matching based on the hydrogen-bonding properties and charged states of nearby receptor atoms.[55]

The scoring grids are also prepared automatically. The grid box dimensions are initially set so that the edges extend 15 Å beyond the matching spheres. The box dimensions are refined to maximize the coverage of the protein without exceeding 2 million grid points at a resolution of three points per angstrom. Polar hydrogens are added to the protein using SYBYL.[48] Four scoring grids are generated: an excluded volume grid using DISTMAP,[56] a united atom AMBER-based van der Waals potential grid using CHEMGRID,[56] an electrostatic potential grid using DelPhi,[57] and a solvent occlusion map using the program SOLVMAP (B Shoichet, unpublished results). The Delphi grid potential is calculated using a dielectric of 2 with the internal low dielectric volume determined by the protein atoms augmented by dummy atoms occupying the

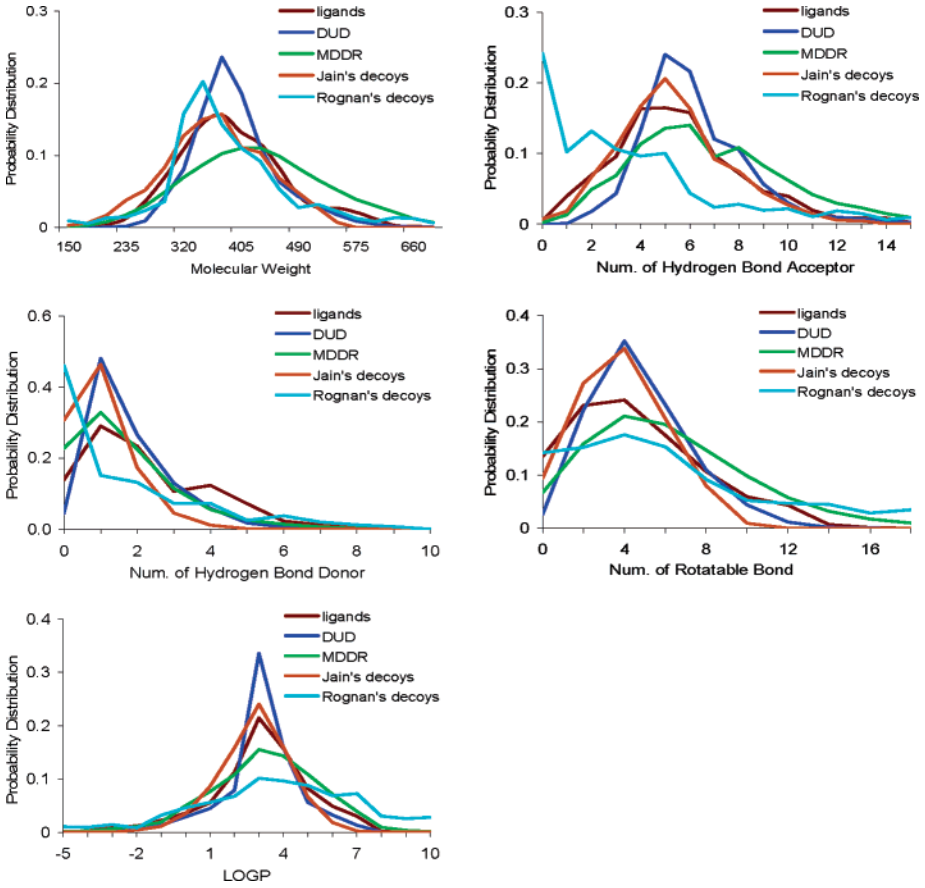


Figure 2. The physical property distributions of the ligands and different sets of decoys. The brown line represents the annotated ligands (2950 compounds); the blue line represents the DUD decoys (95 316 compounds); the green line represents the properties of the MDDR database (98 000 compounds); the orange line represents the Jain's decoys (randomly selected 1000 ZINC druglike compounds), and the cyan line represents the Rognan's decoys (randomly selected 990 ACD compounds).

binding pocket, and an external dielectric of 78 for the external solvent environment. When structural waters, cofactors, or metal ions are present, they are treated as part of the protein.

Docking was performed with DOCK 3.5.54, a flexible-ligand method that uses a force-field-based scoring function composed of van der Waals and electrostatic interaction energies corrected for ligand desolvation.[33,47,56] The sampling of ligand orientations in DOCK3.5.54 can be varied according to several user-defined parameters, which we set to the same values for all 40 systems, as follows. The bin size for both receptor and ligand were set to 0.4 Å and the overlap bin size was set to 0.3 Å. A distance tolerance (dislim) of 1.5 Å was applied for matching the ligand onto the matching spheres, and ligand orientations were rejected if the color of a ligand−receptor pair did not match. For each ligand orientation, the conformational ensemble is filtered for steric complementarity using DISTMAP with the polar and nonpolar close contact limits of 2.3 and 2.6 Å, respectively. Ligand conformations are scored on the basis of the total docking energy ($E_{tot} = E_{ele} + E_{vdw} - \Delta G_{lig-solv}$), which is the sum of electrostatic (E_{ele}) and van der Waals (E_{vdw}) interaction energies, corrected by the partial ligand desolvation energy ($\Delta G_{lig-solv}$).[47] Final energies are computed after 25 steps of rigid-body minimization. A single docking pose with the best total energy score is saved for each docked molecule. For ligands with multiple protonation states and tautomeric forms, only the best scoring representation is retained.

Results

A Directory of Useful Decoys (DUD). We have created DUD as a research tool to benchmark structure-based virtual screening. DUD contains 2950 annotated ligands for 40 diverse targets, plus 36 decoy molecules for each annotated ligand, each decoy having similar physical properties but dissimilar chemical structures to its active counterpart (Table 1). Topological dissimilarities were originally calculated using CACTVS fingerprints, but it is convenient to compare decoys and ligands using Daylight fingerprints, which are more widely used. Of the 95 316 DUD decoys and based on standard Daylight fingerprints,[58] only 90 compounds are above a Tc of 0.85 to any annotated ligand, only 400 compounds are above a Tc of 0.8, and only 1300 are above a Tc of 0.7, indicating that DUD decoys are topologically *dissimilar* to the annotated ligands and are thus likely to be true negatives, although of course we cannot be completely sure that this is the case.[59]

Histograms of five physical properties (molecular weight, number of hydrogen-bond acceptors, number of hydrogen-bond donors, number of rotatable bonds, and log *P*) were calculated for the DUD ligands, DUD decoys, MDDR database compounds, Jain's decoys,[23] and Rognan's decoys[15] (Figure 2). For each of the 40 targets, the DUD decoys are designed to match the physical properties of the specific ligands for that target. Strictly speaking, there is no reason the amalgamation of the 40 decoy sets (the DUD decoys) should provide good decoys for each of the 40 ligand sets (the DUD actives), but in fact they typically do (see Discussion). The properties of the DUD decoys are comparable to the active ligands from which they were generated: they span the same ranges and have maxima at about the same place. Conversely, the uncorrected databases can differ substantially from the physical properties of the DUD ligands (Figure 2). The largest differences are observed for the 1000 decoys introduced by Rognan, which differ substantially in all physical properties from the DUD ligands. The 98 000

MDDR database also differed significantly from the DUD ligands, being typically larger in every physical property (e.g., 9% larger, on average, in molecular weight and having 15% more hydrogen-bond acceptors) than the DUD ligands. The 1000 decoy set introduced by Jain was the second best-matched to the DUD ligands, after only the DUD decoys themselves, typically differing only in being slightly smaller or lower in physical properties (Figure 2). Of course, there is no reason these other decoy sets should match the physical properties of the DUD ligands, since they were not selected to match these ligands. Nor is there any reason to require them to match, unless one believes the DUD ligands to be especially representative of good physical properties, which we do not contend. What the differing physical properties among the decoy sets allow us to probe is how matched and unmatched ligand−decoy sets affect enrichment calculations in docking.

Overall Enrichments. Virtual screening is benchmarked using two criteria: (1) enrichment of annotated ligands among top scoring docked molecules from a database of decoys and (2) the geometric fidelity of the docked poses compared to those of the experimental structures. The docking enrichment factor (EF) reflects the ability of the docking calculations to find true positives throughout the background database compared to random selection. This enrichment factor is calculated as $EF_{subset} = \{ligands_{selected}/N_{subset}\}/\{ligands_{total}/N_{total}\}$.[47] For instance, for a given protein with 100 annotated ligands ($ligands_{total}$) in a database of 98 000 compounds (N_{total}), only one of the known ligands ($ligands_{selected}$) would be expected to be found in any chosen subset of 980 molecules (N_{subset}) by random selection, which corresponds to an enrichment factor of 1. The key results of docking to 40 targets are summarized in Figure 3 and Table 1. Figure 3 shows the overall profile of percentage of ligands found (*y*-axis) plotted as a function of the percentage of the ranked docked database (*x*-axis in logarithmic scale) for each system. Here, we present two different enrichments using two different background databases, namely, the "entire database" contains the 2950 DUD ligands and the 95 316 DUD decoys (blue line), while the "own decoys" only includes the native ligands and their corresponding decoys (red line). The percentage of true ligands found by docking at any given percentage of the docking ranked database should always be greater compared to being chosen by random selection (gray line). The higher the percentage of known ligands found at a given percentage of the ranked database, the better the enrichment performance of the virtual screening. In general, the docking enrichments are poorer against the "own decoys" than against the entire database; for some targets, the enrichment difference between the entire database and the "own decoys" is dramatic (e.g. TK, PNP, and SAHH).

The enrichment results for entire-database docking may be summarized using three enrichment indicators: EF_{max} (maximum enrichment factor), EF_1 (enrichment factor at 1% of the ranked database), and EF_{20} (enrichment factor at 20% of the database) (Table 1). EF_{max} and EF_1 present the early enrichment, while EF_{20} presents the late-stage database screening. Significant enrichment is obtained for most targets, with an average EF_{max} of 49.1, where 20 systems have EF_{max} greater than or very close to 30, 10 systems have an EF_{max} less than 5, and only 4 out of these 10 systems fail to enrich their native ligands above random. Superior enrichments are observed in seven diverse targets with EF_{max} greater than 100. On average, 17.3% and 52.9% of the known ligands can be found in the top 1% and 20% of the docking ranked database, respectively, corresponding to enrichment factors of 17.3 and 2.6. It is also notable that six out of 10 targets with poor enrichment are kinases. We now take up in more detail the dependence on decoys, docking specificity via cross docking, and consider six representative targets.

1. Enrichments against DUD Compared to Uncorrected Databases. To test the influence of decoys on docking performance, we docked exactly the same ligand sets against 12 different targets, varying only the background decoy database. We compared the 98 000 compounds of the MDDR, the 98 266 compounds of DUD, and the 1000 compound sets introduced by Rognan[15] and Jain.[23] For comparing the performance of the large DUD and MDDR databases to the smaller Rognan and Jain sets, we used receiver operator characteristic (ROC) curves to avoid biases introduced in enrichment plots when the ratio of actives to decoys grows large.[60] ROC curves plot sensitivity (Se) and specificity (Sp), where $Se_{subset} = \{ligands_{selected}/ligands_{total}\}$ and $Sp_{subset} = \{(decoys_{total} - decoys_{selected})/decoys_{total}\}$. We plotted the ROC curves as (1 − Sp) (i.e., % selected decoys) versus Se (i.e., % selected ligands) (Figure 4). Like an enrichment plot, the further away the ROC curve is above the diagonal, the better the docking enrichment. Docking enrichments typically followed the following trend: the Rognan decoys led to the best enrichments, followed closely by the MDDR decoys, then the Jain decoys, with the worst enrichments against DUD. Here, better enrichment means only less competitive decoys. Thus, targets that had poor or no enrichment using DUD had very respectable enrichments against the other decoy sets, using exactly the same ligands and docking protocols. To investigate whether a smaller decoy database itself introduces artificial enrichment and thus unfairly biases the smaller decoy sets, ROC curves were also generated using a randomly selected 1000 compounds from DUD and compared with ROC curves using the entire DUD. No difference was observed (not shown), suggesting that there is little size-dependent behavior using the entire DUD versus a random portion of DUD. The more competitive behavior of the DUD decoys presumably reflects their closer physical similarity to the ligands docked; indeed, the monotonic order of enrichments follows the level of dissimilarity of the decoys to the ligands, with the Rognan decoys being the most dissimilar and the Jain and DUD decoys being the most similar and correspondingly leading to the worst (i.e. most competitive) enrichments.

2. Docking Specificity via Cross-Docking Simulations. We docked DUD against all 40 targets and compared the enrichment of each ligand set against each target (Table 2). We used two enrichment indicators, ET_{max} and ET_{20}, to define the enrichment performance for each matrix unit as very good ($ET_{max} \geq 30$ and $ET_{20} \geq 3$), good ($30 > ET_{max} \geq 20$ and $3 > ET_{20} \geq 2.5$), medium ($20 > ET_{max} \geq 10$ and $2.5 > ET_{20} \geq 2$), or poor ($ET_{max} < 10$ and $ET_{20} < 2$). An exception (e.g. $ER_{antagonist}$) is made when one of the two enrichment indicators is well above its defined cutoff while the other is marginally below its cutoff. In these cases an averaged enrichment performance classification is assigned. Several features of the cross-docking table are noteworthy. First, it is a sparse matrix, mostly white, showing that most annotated ligand sets are not highly enriched against most targets. Second, many of the diagonal elements are black or red, indicating very good or good enrichment of the target's own ligands. Third, many of the off diagonals make sense. For example, serine protease ligands (thrombin, trypsin and factor Xa) are enriched against other serine protease targets; nuclear hormone receptor ligands are enriched against most nuclear hormone receptors, such as androgen receptor (AR), mineralocorticoid receptor (MR), and estrogen receptor (ER); and nucleoside analogues are enriched against most of the nucleo-

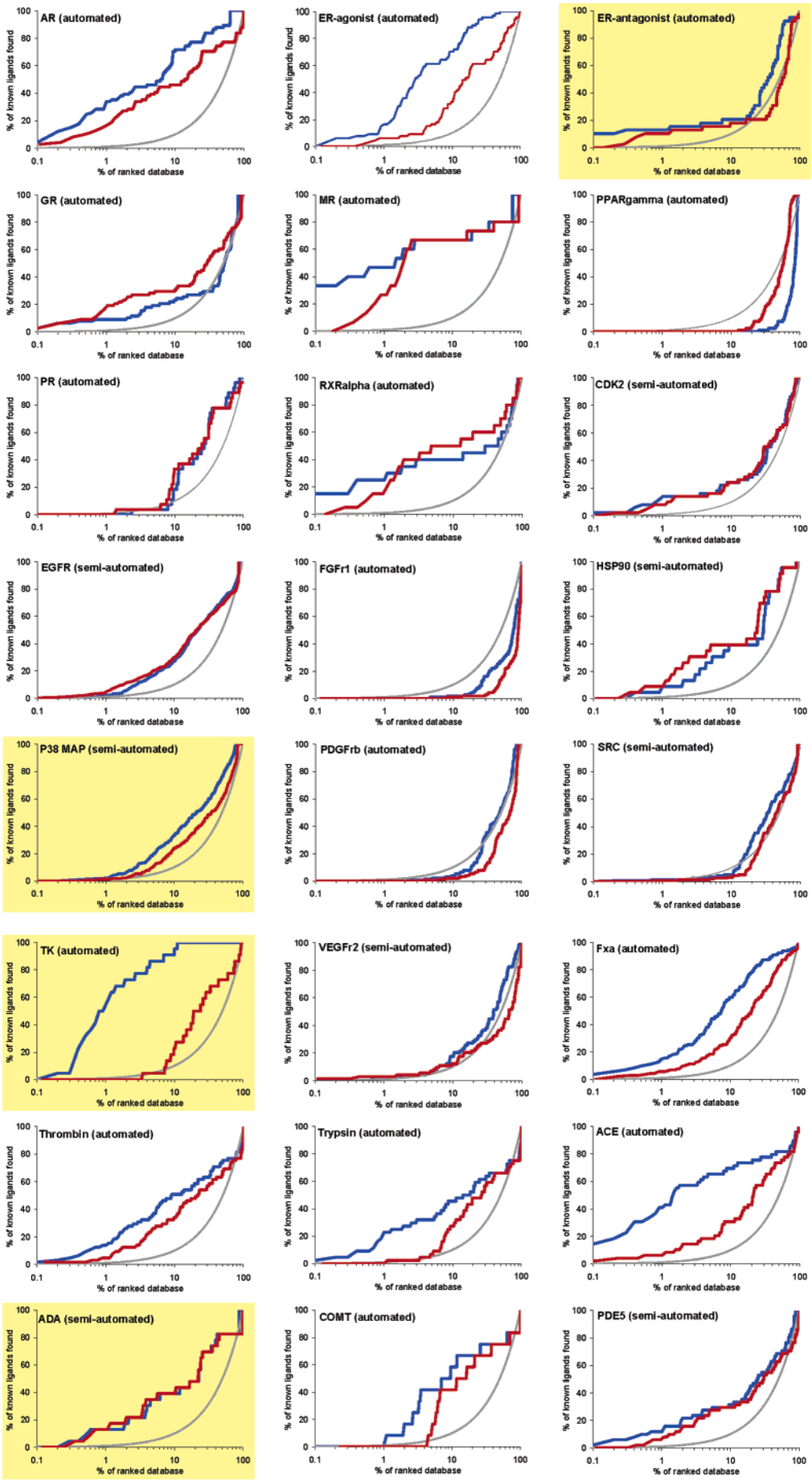


Figure 3. (Continued on next page).

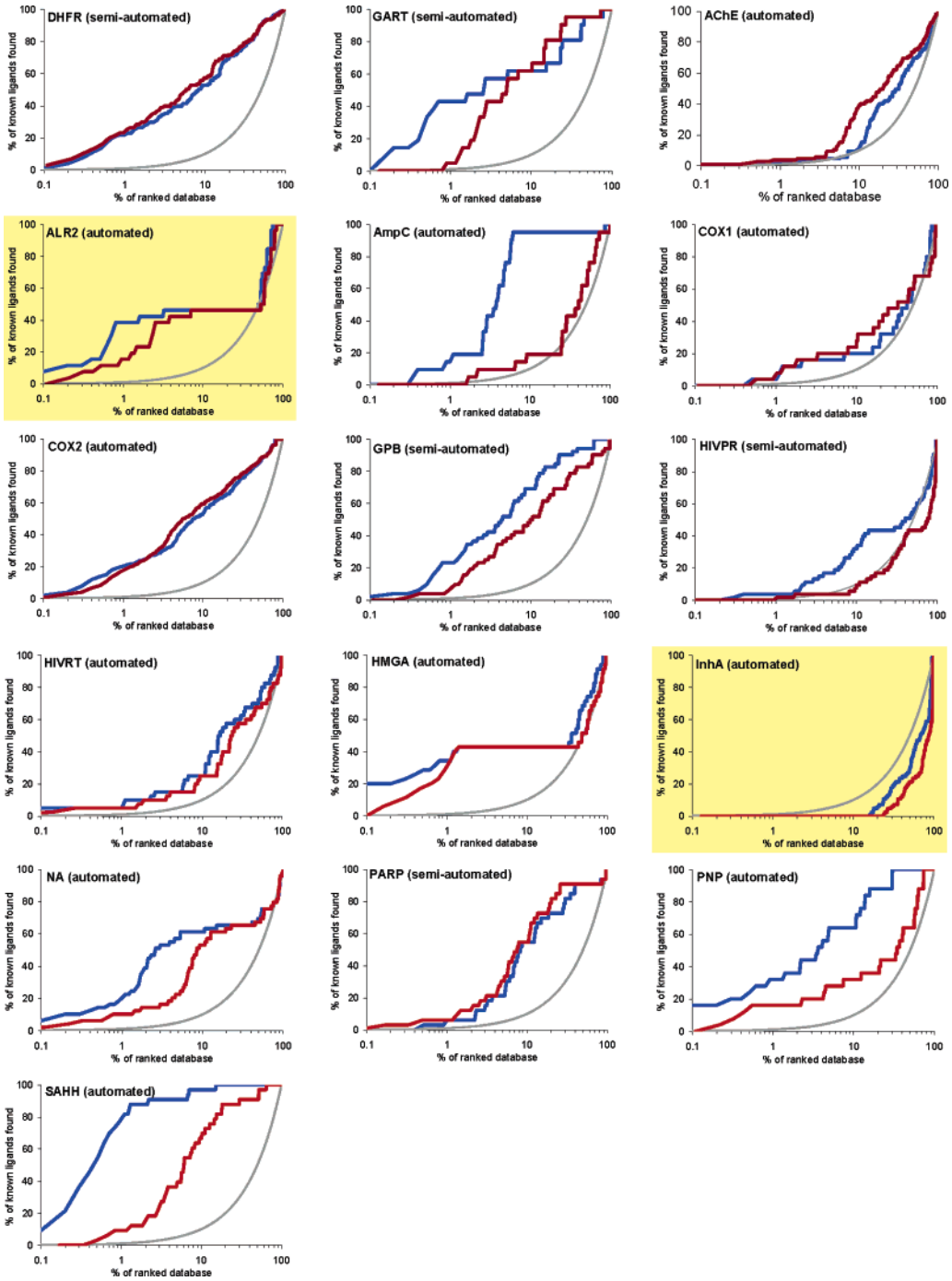


Figure 3. Docking enrichment plots for 40 protein targets using DUD. The docking ranked database (*x*-axis) is plotted against the percentage of known ligands found by calculations (*y*-axis) at any given percentage of ranked database. Targets are listed in same order as in Table 1, and six representative systems are highlighted in light yellow (see the text). The gray line represents the results expected from selecting ligands randomly; the blue line is docking enrichment against the entire DUD database (98 266 compounds), and the red line is the docking enrichment against the "own decoy" subset for any target. "Automated" represents the results achieved from the fully automated procedure; "semiautomated" represents the enrichments obtained with some expert intervention.

side-recognizing enzymes, such as thymidine kinase (TK), purine nucleoside phosphorylase (PNP), and S-adenosyl-homocysteine hydrolase (SAHH). Fourth, the kinases generally show poor enrichments, not just of their own ligands but of all ligand lists. More generally, we note that when the cognate ligands are not enriched against a target, other ligand lists are also not enriched, giving rise to blank rows. Conversely, when its own ligands are well-enriched against a target, other ligand lists are often also enriched.

3. Automated vs Semiautomated docking. The large number of docking targets motivated us to develop a fully automated docking engine. We were able to automate most of the steps formerly performed manually, resulting in satisfactory enrichments for 24 of 40 targets. For the remaining 16 targets, we resorted to a semiautomated procedure involving expert intervention in the preparation of the receptor binding site. Thirteen of the 16 targets with poor enrichment using the automated protocol were improved with expert intervention (Supporting Information, Figure S4). This intervention was often trivial; we did not try very hard to maximize the docking performance and suspect that further intervention could improve the results even more, but that was not the goal of this study.

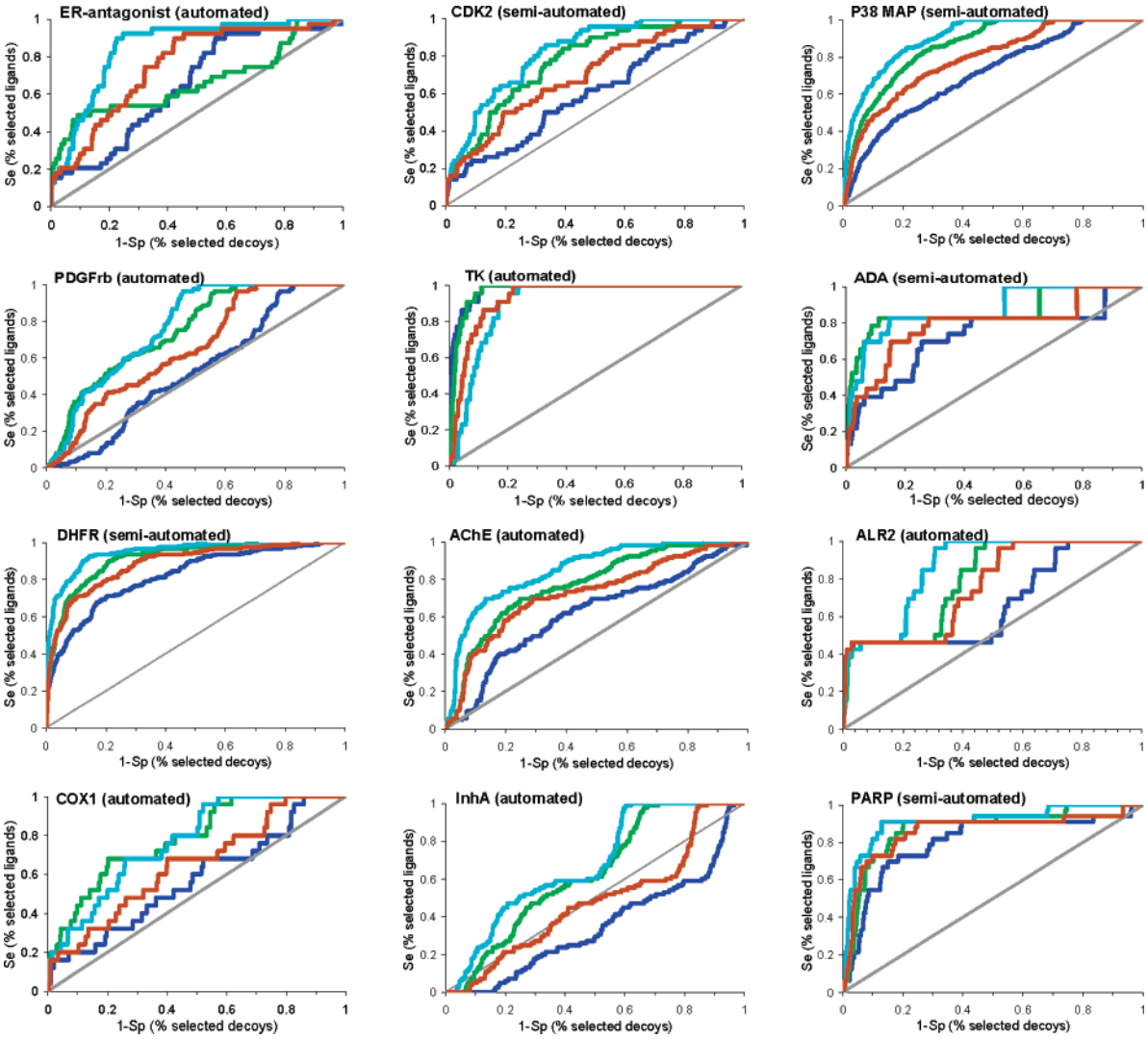


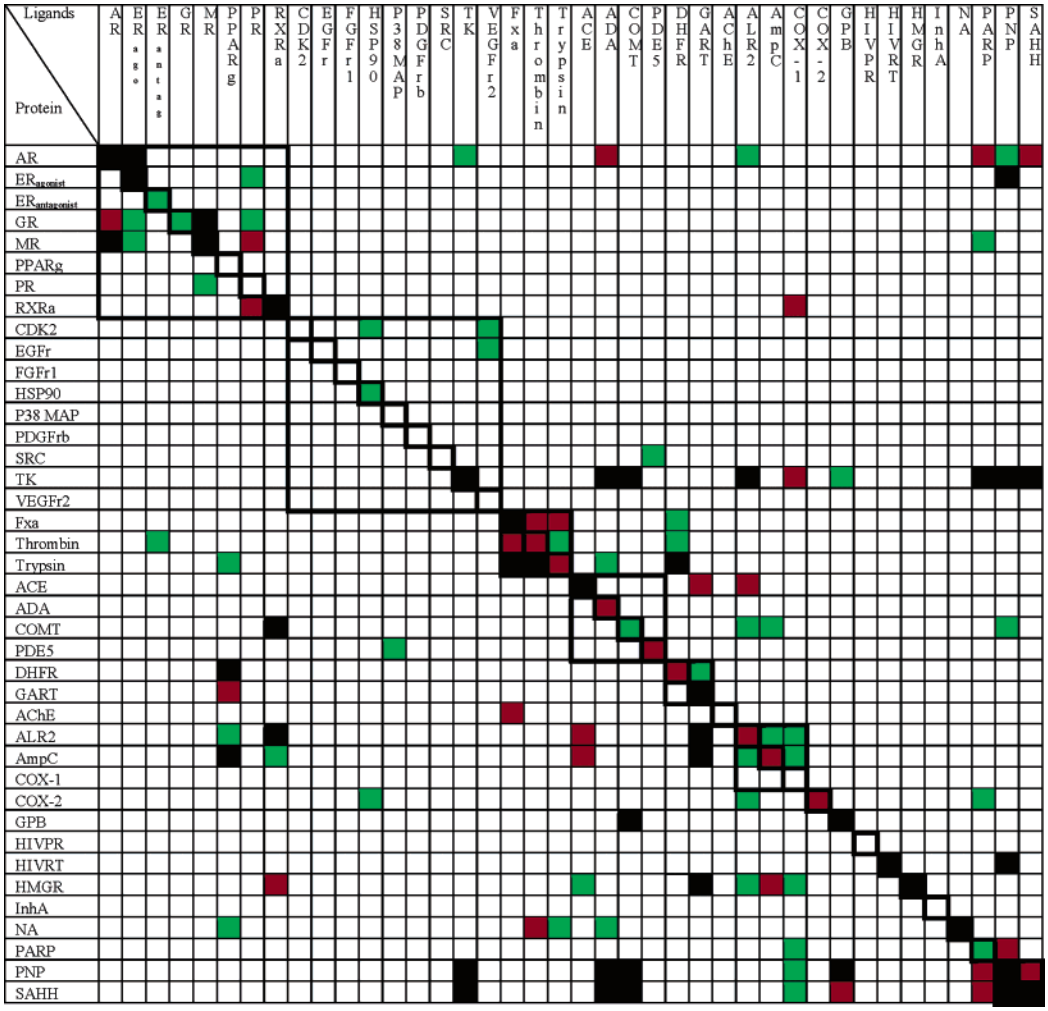
Figure 4. ROC curves for 12 targets using four different background databases, DUD (blue), MDDR (green), Jain's decoys (orange), and Rognan's decoys (cyan). The gray line represents the results expected from random selection of ligands. The ROC curves were plotted as the Se (% selected actives) versus (1 − Sp) (% selected decoys). The same annotated ligands were used for the different background databases. Targets are listed in the same order as in Table 1.

4. Detailed Results for Six Representative Systems. We examined several representative targets in greater detail (highlighted in Table 1 and Figure 3). ER and TK were chosen on the basis of their strong ligand enrichment as well as a substantial number of published docking studies.[15,19,20,23,28−30,37−40] P38 MAP kinase was chosen to represent poorly performing protein kinases. ADA was chosen to represent targets that failed with the fully automated docking engine, but were rescued by the semiautomated procedure. ALR2 was chosen to represent targets with intermediate enrichment. InhA was chosen to represent what we consider a failure of our docking method. For these six representative systems, the docking accuracy is presented by both the enrichment performance and the ability to reproduce the binding geometry observed in the crystallographic complex structure. In assessing the docked binding geometries, we only consider the geometry of the crystallographic ligand produced as part of the overall DUD database screen without optimization; the ligand shown has been prepared as every other DUD molecule, starting from the SMILES string representation, to avoid bias. Depending on the size of the binding pocket and our sampling criteria, docking DUD takes several hours to several days per target on a single 2.8 GHz CPU (Table 3).

4A. Estrogen Receptor ($ER_{antagonist}$). ER is considered to be an easy docking target because of its deeply buried hydrophobic binding site and high-affinity ligands.[15] Consistent with this view, automated docking achieves a significant early enrichment with an EF_{max} of 101.6 in the top 0.1% of the ranked database. This corresponds to finding four ER antagonists among the top scoring 98 compounds from a screen of 98 000 compounds. From automated docking, the docked pose of the crystallographic ligand, included and prepared as part of DUD, correctly reproduces the crystallographic binding pose of 4-hydroxytamoxifen (Figure 5A).

As was observed for almost all targets, docking enrichments were strongly influenced by the choice of background database. For estrogen receptor, the three other decoys sets (the 990 Rognan decoys, the 1000 Jain decoys, and the 98 000 MDDR decoys) all led to much better enrichments than did DUD, with the early enrichment being particularly striking (Figure 4). Since the docking parameters and the ligands were exactly the same in each calculation, the better enrichments compared to the DUD results can only mean that the other databases present easier decoys. This view is substantiated by comparing the physical properties of the ER antagonists to the decoys in each background database. In DUD, these properties are closely

Table 2. Matrix of Cross-Enrichments[a]



[a] The color-coded table unit presents poor (white), medium (green), good (red), and very good enrichment (black). Dark boxes are drawn around related targets (nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes). Very good ($ET_{max} \geq 30$ and $ET_{20} \geq 3$), good ($30 > ET_{max} \geq 20$ and $3 > ET_{20} \geq 2.5$), medium ($20 > ET_{max} \geq 10$ and $2.5 > ET_{20} \geq 2$), and poor ($ET_{max} < 10$ and $ET_{20} < 2$). The only exception is when one of the two enrichment indicators is well above the defined cutoff while the other is marginally below the defined cutoff, and then an averaged enrichment performance is assigned to compensate it.

Table 3. Docking Statistics on Six Representative Targets

receptor	unique molecules scored[a]	total molecules scored[b]	orientations sampled per molecule	conformations sampled per molecule	total configurations scored[b]	total time (h)[c]
ER	97 427	416 990	1 895	6 543	2.69×10^{10}	54.4
P38 MAP	93 887	294 917	592	7 875	8.97×10^{9}	20.1
TK	37 240	180 451	3 437	4 302	2.67×10^{9}	21.9
ADE	85 053	297 400	14 632	5 308	2.19×10^{10}	65.5
ALR2	98 724	430 313	4 272	10 109	1.44×10^{11}	296.4
InhA	97 668	429 579	2 325	6 809	5.87×10^{10}	123.5

[a] Only orientations and configurations passing the steric filter were scored. [b] Some molecules were represented in the database in multiple rigid fragment, protonation, and tautomeric forms. [c] Scaled to reflect time on a 2800-MHz Pentium IV.

matched, by design. In the other decoys sets, there are substantial differences. For example, the molecular weight varies from 380 to 460 for 90% of ER antagonists. Sixty percent of DUD compounds are within this range, whereas only 40%, 42%, and 44% of compounds in Rognan's set, MDDR, and Jain's set satisfy this criterion, respectively. In addition, the ER binding site requires specific hydrogen-bonding interactions, whereas a large portion of Rognan's decoys lack hydrogen-bonding functionalities, further increasing the likelihood of them being easier decoys than either the MDDR or Jain's decoys.

4B. Thymidine Kinase (TK). TK is considered to be a difficult target for docking because of receptor flexibility, a highly exposed binding pocket, the importance of water-bridged interactions, and the low affinity of most ligands.[15] Despite these drawbacks, a high overall enrichment was achieved for this target (Figure 3). From automated docking, the docked pose corresponds closely to the crystallographic structure (Figure 5B). We note that TK is among the most promiscuous targets, in that it also highly enriches non-native ligands (Table 2), where those nonligands are either nucleoside analogues (PNP and SAHH ligands) or highly polar or charged small compounds (ADA, COMT, ALR2, and PARP ligands).

Of the 12 targets where we compared decoy sets, TK was the only one where DUD led to better enrichments than the

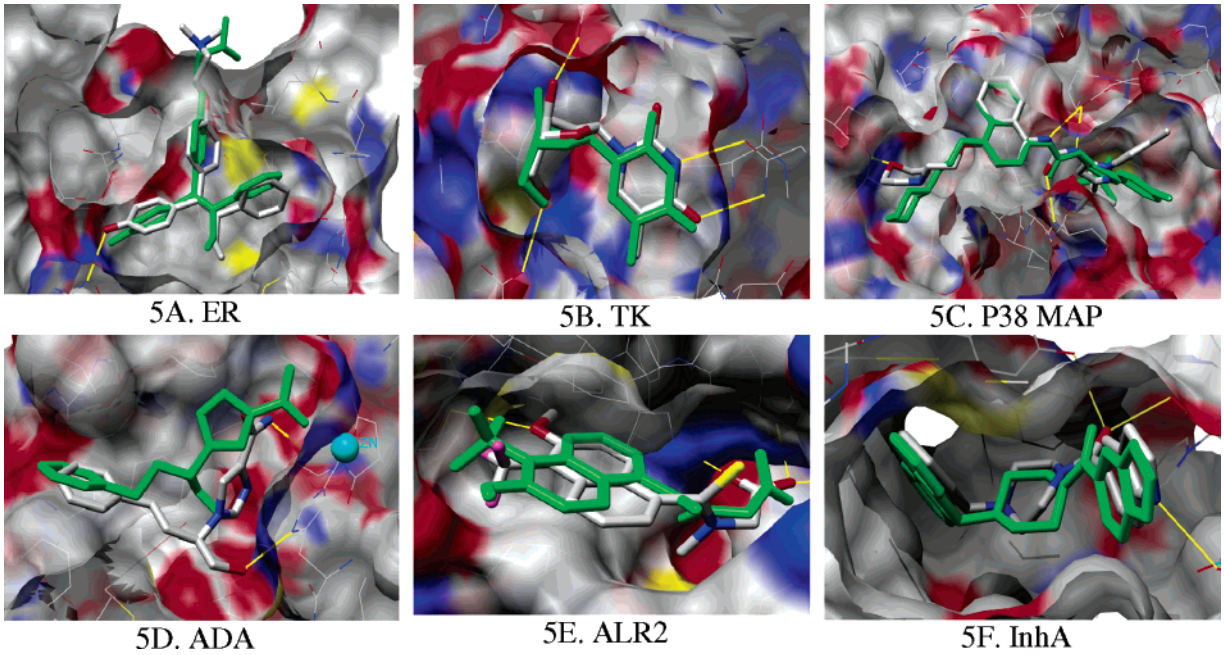


Figure 5. Binding pose predictions for docked ligands (green) superposed on crystallographic structures (colored by atom type) for six representative targets. Key hydrogen bonds are shown by yellow lines, and the protein molecular surface[65] is colored by atom type. Images generated with Chimera.[66]

other sets. The TK inhibitors are much smaller and more hydrophilic than most other ligand sets and correspondingly smaller than much of DUD. The molecular weight of 70% of the TK inhibitors is between 210 and 290, but only a small portion of the background database compounds fell within this range, 5%, 9%, 10%, and 19% of DUD, Jain's set, MDDR, and Rognan's decoys, respectively. The idiosyncratic nature of the TK ligands is brought home by comparing the very good enrichment against all of DUD compared with the relatively miserable enrichment compared to the "own decoys" of the TK ligands (Figure 3, highlighted). This is one of the rare cases where the overall DUD decoys had very different physical properties from the "own decoys" of a particular target.

4C. P38 Mitogen Activated Protein Kinase (P38 MAP). P38 MAP kinase is a challenging target for docking due to its structural flexibility. For many ligands, a new allosteric site is induced upon binding.[61] Docking is also complicated by the high degree of solvent exposure and a relatively shallow, hydrophobic cleft; taken together, these features result in poor enrichment and docking geometries that miss critical interactions. Many highly ranked decoys explore irrelevant binding regions, and manual intervention had little effect. From the automated docking, the docked ligand reproduces the correct binding geometry of urea and naphthyl groups, but not the morpholino substituent on the naphthyl ring, which are thought to be crucial for potency[61] (Figure 5C). Similar defects were observed with other protein kinases (PDGFrb, VEGFr2, EGFr, SRC, and FRFr1).

Like many other targets, the enrichment performance in decreasing order is Rognan's set, MDDR, Jain's set, and DUD. As for TK, a molecular size dependence biases the docking enrichments. The molecular weight ranges from 320 to 410 for 75% of the P38 kinase inhibitors, whereas for DUD, Jain's set, and MDDR, 67%, 54%, and 37% of molecules fall within this range, respectively, directly corresponding to their enrichment performance. Although sixty percent of Rognan's set falls within this weight range, a lack of suitable hydrogen-bonding functional groups in the set ensures the best enrichment performance.

4D. Adenosine Deaminase (ADA). ADA is one of four metalloenzymes we targeted. Its binding pocket is large and contains a zinc ion coordinated by three histidines. No enrichment is achieved for this target via fully automated docking. After manually redistributing the partial atomic charges of the N^{ϵ}-atoms in the ligating histidine residues to the Zn ion, which we have previously found to be important for docking to metalloenzymes,[34] and included one structural water in the active site, the enrichment improved significantly (Supporting Information, Figure S4). From automated docking, the docked ligand is approximately matched with the crystallographic ligand. Although some key groups and interactions have shifted, the ligand still occupies the same space as the crystallographic ligand, inhibiting approach to the catalytic zinc (Figure 5D). Electrostatics plays an important role in these metalloenzymes. All ADA inhibitors contain hydrogen-bond donors ranging from 1 to 5, whereas 95%, 75%, 69%, and 45% for DUD, Jain's set, MDDR, and Rognan's decoys do, respectively. Correspondingly, the enrichments with the MDDR and Rognan decoys are better than those with DUD and Jain's decoys.

4E. Aldose Reductase (ALR2). ALR2 presents a solvent-exposed binding surface that requires both good polar and hydrophobic complementarity between the enzyme and inhibitors. Intermediate enrichment performance is observed in ALR2 with a good early enrichment ($EF_1 = 38$) but a rather weak enrichment at a later stage of database screening ($EF_{20} = 2.3$), which might be attributed to the conformational changes induced by the binding of different sizes of ligands. From automated docking, the docking pose reproduces the critical polar interactions within the protein binding site and overlaps with the crystallographic binding pose except for the flipping of the aromatic ring (Figure 5E). Unsurprisingly, both the molecular size and hydrogen-bonding capacity of the decoys influence docking enrichment. The molecular weight ranges from 260 to 400 for 85% of the ALR2 inhibitors, whereas 72%, 68%, 66%, and 45% of the DUD decoys, Jain's set, Rognan's decoys, and the MDDR compounds do, respectively, which explains the better enrichment using MDDR. Although, molecular size alone

is not adequate to distinguish Rognan's set from the others, the absence of suitable hydrogen-bonding functional groups ensures its best enrichment performance.

4F. InhA. The worst docking scenario occurs with InhA (Figure 3), where docking gives no enrichment despite reasonable docking geometries of the crystallographic ligand and related analogues (Figure 5F), a reminder that reproducing the crystallographic binding pose is a necessary but not sufficient criterion for evaluating virtual screening. The docking enrichment decreases in the order: Rognan's set, MDDR, Jain's set, the DUD decoys. It seems that the requirements of the presence of hydrogen-bond acceptors and the steric fitness to the InhA binding site are important for ligand binding, which makes the DUD decoys the most competitive.

Considering all 40 targets, docking enrichment varied from excellent to poor, returning what we consider to be satisfactory but not stellar enrichments. Typically, poor enrichments could be attributed to sampling of ligand conformations and, in the cases of kinases, undersampling those of the receptor. There were other targets for which the differentiation between ligands and strong decoys was uninspiring, reflecting failures of our physics-based scoring function. These are all important weaknesses in our docking program and are not uncommon in the field. As important as they are, they are not the subject of this paper, which is focused on database bias and its role in evaluating docking enrichment factors.

Discussion

Three key results emerge from this study. Perhaps the one that will have the greatest pragmatic impact is the creation of the DUD database itself. DUD is composed 2950 annotated actives together with compounds having dissimilar topology but similar physical properties to the active ligands. Because of these properties, it provides a challenging, but relatively unbiased, metric for evaluating docking performance. As DUD is composed entirely of compounds in the public domain, it may be used without restriction; it may be downloaded from http://blaster.docking.org/dud/. Second, by docking all ligand sets against all 40 targets, we unintentionally undertook a very large "cross-docking" experiment. The specificity matrix that results from these cross-docking results suggests interesting patterns relating to docking promiscuity and level of difficulty for docking targets. Finally, it is somewhat surprising that a fully automated docking pipeline yielded reasonably good results for 24 of 40 cases and that minor expert intervention improved docking enrichments for 13 of the remaining 16 targets.

DUD is by far the largest and most comprehensive public data set for benchmarking virtual screening programs of which we are aware. The forty targets used to create DUD offer a diverse range of binding site types: some have deeply buried hydrophobic pockets, such as estrogen receptor and COX-2; some have more open binding sites displaying both polar and apolar binding regions, such as DHFR and P38 MAP kinase; and some have highly solvent-exposed polar sites, such as thymidine kinase and neuraminidase. The receptor diversity is reflected in the ligands they recognize: some are mostly hydrophobic (e.g. the ER ligands have log *P* values in the range of 3−8), some are highly polar (e.g. the TK ligands have log *P* values between −3 and 2), some are mostly cationic (e.g. the thrombin ligands have one or two positively charged groups) and some are mostly anionic (e.g. the GART ligands typically have two negatively charged groups). This binding site diversity allows us to evaluate the robustness and generality of our docking methods with some confidence that the range of targets is representative.

Our results are consistent with the observation by Verdonk[27] that enrichment depends on the background database used and suggest that misleadingly good enrichment may result if the physical properties of the decoys are easily distinguished from those of the actives (Figure 4). It is for this reason that the uncorrected databases, using exactly the same basis ligands, consistently lead to better enrichments than docking with DUD. For this purpose, good enrichments indicate nothing other than relatively poor decoys. Even so large and diverse a database as the MDDR led to artificial improvements in enrichment factors, typically by half a log over DUD. This is not to say that DUD itself is ideal. An indication of this is the greater stringency often provided by the "own decoys", i.e., those decoys matched only to the annotated ligands for a particular target, compared to DUD overall (compare red and blue curves in Figure 3). Indeed, it could be argued that it is the "own decoys" that should always be used and not the amalgamated DUD. Our own view is that both "own decoy" and amalgamated DUD docking should be performed, because they present distinct challenges to the docking program. We are aware, also, that the five physical properties we used to match ligands with decoys were not comprehensive, and it is likely that other properties could usefully be included. DUD can be seen as a procedure for building a decoy database as much as an instance of one. Thus, there is probably no such thing as a perfect single decoy set for testing docking algorithms against all targets, but there certainly are better and worse ones, and the former offer better protection against artifactual performance to the unwary docker.

In docking each of 40 DUD ligand and decoy sets against all 40 targets, we unintentionally undertook a very large cross-docking experiment and so arrived at a measure of library-scale specificity. "Cross-docking" typically investigates the specificity of a particular ligand for a particular protein conformation[26,62,63] or, more rarely, the specificity of a particular ligand for a few possible targets.[64] It is a more subtle gauge of docking success than simply the distance to a crystallographic orientation. Correspondingly, the specificity of ligands for their cognate receptor versus the other 39 targets more stringently measures docking success than enrichment against a single target alone. Most of the diagonal elements in the cross-docking matrix indicate decent to very good enrichment (Table 2), and the matrix overall is sparse, with little enrichment against off-diagonal targets. Specificity was rarely perfect, however. Often, off-diagonal promiscuity reflected similarities among the targets (squared regions along the diagonal in Table 2). Thus the nuclear hormone analogues have moderate to very good enrichments against several noncognate nuclear hormone receptors and the serine protease inhibitors often do well against not only their own targets, but also against several of the other serine proteases that recognize similar functionality. But even this is not the full story—many targets that enriched their cognate ligands also enriched ligands of unrelated targets, typically those with similar physical properties as their own ligands. Indeed, one of the most striking features of the cross-docking matrix is that targets that had very good enrichments for their cognate ligands typically also had good enrichments against a few other ligand sets, whereas targets that had poor enrichment for their own ligands typically had no enrichment against any other sets (white rows in Table 2). These latter, white-row targets are effectively difficult targets, at least for our docking program.

Somewhat to our surprise, fully automated docking performed well in 24 of 40 cases (Figure 3, Table 1). Typically in docking, one spends a great deal of time visually inspecting the receptor site, identifying binding site hot spots, adjusting the protonation

states and orientation of rotatable protons of critical binding site residues, and deciding which structural waters, cofactors, or metal ions should be included in the model. This artisanal treatment becomes less feasible in larger studies, such as this one, and indeed has inhibited the proteome-level efforts in docking that are common in related fields, such as comparative modeling of protein structures.[49] Whereas we do not anticipate the end of artisanship in docking campaigns, after all we had to return to manual intervention in 40% of the targets, the relative success of the automated pipeline suggests that efforts to expand such treatments merit more attention.

In docking screens, one cannot expect to correctly predict binding affinity or even monotonically rank order the ligands, so the method falls back on the weak metric of ligand enrichment. Enrichment is a weak measure of docking success because it is always measured relative to the decoys in the database. The very same docking program with the same ligands can have very good or very mediocre enrichments with worse or better decoys. This is currently something we must live with as a field. What we can do is minimize the biases inherent in enrichment factors by matching the physical properties in decoys with those of the ligands. DUD attempts to do just this, and may provide a useful benchmarking set for the field. It is available to all investigators at http://blaster.docking.org/dud/.

Acknowledgment. Supported by NIH grants GM71896 (to B.K.S. and J.J.I.) and GM59957 (to B.K.S.). We thank MDL Inc. for the MDDR database and ISIS software, Schrodinger Inc for LigPrep, Xemistry GmbH for CACTVS, OpenEye (Santa Fe, NM) for OEChem, Molecular Networks GmbH for Corina, Daylight for the fingerprint and smiles toolkits. We thank Austin Kirchner and David Lorber for scripts, code, and assistance in the early stages of this work, and Kaushik Raha, Alan Graves, and Michael Mysinger for reading the manuscript.

Supporting Information Available: Schematic description of the automated docking pipeline, selected property distribution historgrams and 2D depictions of typical molecules for annotated ligands and their decoys, property distribution histograms of Rognan's ligands and decoys, enrichment plots comparing the semiautomated with the fully automated docking procedure, and complete listings of the modified parameter files used. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) DesJarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; et al. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 6644−6648.
(2) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-based discovery of inhibitors of thymidylate synthase. *Science* **1993**, *259*, 1445−1450.
(3) Li, S.; Gao, J.; Satoh, T.; Friedman, T. M.; Edling, A. E.; et al. A computer screening approach to immunoglobulin superfamily structures and interactions: Discovery of small non-peptidic CD4 inhibitors as novel immunotherapeutics. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 73−78.
(4) Gruneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588−3602.
(5) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure (Camb)* **2002**, *10*, 1013−1023.
(6) Huang, N.; Nagarsekar, A.; Xia, G.; Hayashi, J.; MacKerell, A. D., Jr. Identification of non-phosphate-containing small molecular weight inhibitors of the tyrosine kinase p56 Lck SH2 somain via in silico screening against the pY + 3 binding site. *J. Med. Chem.* **2004**, *47*, 3502−3511.
(7) Song, H.; Wang, R.; Wang, S.; Lin, J. A low-molecular-weight compound discovered through virtual database screening inhibits Stat3 function in breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102, 4700−4705.
(8) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(9) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.
(10) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(11) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 1−6.
(12) 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 Discov* **2004**, *3*, 935−949.
(13) Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. Docking: Successes and challenges. *Curr. Pharm. Des.* **2005**, *11*, 323−333.
(14) 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.
(15) 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.
(16) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, 44, 1035−1042.
(17) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein−ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.
(18) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(19) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(20) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; et al. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, 47, 1750−1759.
(21) 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.
(22) 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.
(23) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2005**.
(24) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.
(25) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.
(26) Velec, H. F.; Gohlke, H.; Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.
(27) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; et al. Virtual screening using protein−ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.
(28) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(29) Yang, J. M.; Shen, T. W. A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **2005**, *59*, 205−220.
(30) Li, H.; Li, C.; Gui, C.; Luo, X.; Chen, K.; et al. GAsDock: A new approach for rapid flexible docking based on an improved multi-population genetic algorithm. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4671−4676.
(31) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.
(32) Diller, D. J.; Li, R. Kinases, homology models, and high throughput docking. *J. Med. Chem.* **2003**, *46*, 4638−4647.
(33) Lorber, D. M.; Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* **2005**, *5*, 739−749.
(34) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* **2005**, *44*, 12316−12328.
(35) Irwin, J. J.; Shoichet, B. K. ZINC−A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; et al. The Protein Data Bank. *Nucleic Acid. Res.* **2000**, *28*, 235−242.
(37) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46*, 3045−3059.
(38) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781−5789.
(39) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: Evaluation of current docking tools. *J. Mol. Model. (Online)* **2003**, *9*, 47−57.
(40) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.
(41) Claussen, H.; Gastreich, M.; Apelt, V.; Greene, J.; Hindle, S. A.; et al. The FlexX database docking environment−Rational extraction of receptor based pharmacophores. *Curr. Drug Discovery Technol.* **2004**, *1*, 49−60.
(42) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Des* **2004**, *18*, 333−344.
(43) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076−5084.
(44) Mozziconacci, J. C.; Arnoult, E.; Bernard, P.; Do, Q. T.; Marot, C.; et al. Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, 48, 1055−1068.
(45) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.
(46) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(47) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(48) *SYBYL*; 6.7 ed.; Tripos Associates: St. Louis, MO.
(49) Eswar, N.; John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V. A.; et al. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **2003**, *31*, 3375−3380.
(50) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; et al. A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004**, *55*, 351−367.
(51) Brenk, R.; Irwin, J. J.; Shoichet, B. K. Here Be Dragons: Docking and Screening in an Uncharted Region of Chemical Space. *J. Biomol. Screen* **2005**, *10*, 667−674.
(52) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709−713.
(53) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. The MIDAS display system. *J. Mol. Graph.* **1988**, *6*, 13−27.
(54) 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.
(55) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.
(56) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(57) Nicholls, A.; Honig, B. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson−Boltzmann equation. *J. Comput. Chem.* **1991**, *12*, 435−445.
(58) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; J. Daylight CIS Inc., April 17, 2006. Retrieved October 19, 2006 from the World Wide Web: http://www.daylight.com/dayhtml/doc/theory/index.html.
(59) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(60) 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.
(61) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.
(62) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377−395.
(63) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.
(64) Sotriffer, C. A.; Dramburg, I. "In situ cross-docking" to simultaneously address multiple targets. *J. Med. Chem.* **2005**, *48*, 3122−3125.
(65) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305−320.
(66) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; et al. UCSF Chimera−A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(67) Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; et al. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* **2004**, *28*, 401−407.
(68) Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; et al. Structure−activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **2001**, *14*, 280−294.
(69) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(70) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.
(71) Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; et al. Structure−activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **2004**, *11*, 775−785.
(72) Dymock, B. W.; Barril, X.; Brough, P. A.; Cansfield, J. E.; Massey, A.; et al. Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J. Med. Chem.* **2005**, *48*, 4212−4215.
(73) Hennequin, L. F.; Thomas, A. P.; Johnstone, C.; Stokes, E. S.; Pl inverted question marke, P. A.; et al. Design and structure−activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors. *J. Med. Chem.* **1999**, *42*, 5369−5389.
(74) Hennequin, L. F.; Stokes, E. S.; Thomas, A. P.; Johnstone, C.; Ple, P. A.; et al. Novel 4-anilinoquinazolines with C-7 basic side chains: Design and structure activity relationship of a series of potent, orally active, VEGF receptor tyrosine kinase inhibitors. *J. Med. Chem.* **2002**, *45*, 1300−1312.
(75) Sun, L.; Tran, N.; Liang, C.; Tang, F.; Rice, A.; et al. Design, synthesis, and evaluations of substituted 3-[(3- or 4-carboxyethylpyrrol-2-yl)methylidenyl]indolin-2-ones as inhibitors of VEGF, FGF, and PDGF receptor tyrosine kinases. *J. Med. Chem.* **1999**, *42*, 5120−5130.
(76) Bohm, M.; St rzebecher, J.; Klebe, G. Three-dimensional quantitative structure−activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.
(77) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure−activity relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.
(78) Varney, M. D.; Palmer, C. L.; Romines, W. H., 3rd; Boritzki, T.; Margosiak, S. A.; et al. Protein structure-based design, synthesis, and biological evaluation of 5-thia-2,6-diamino-4(3*H*)-oxopyrimidines: Potent inhibitors of glycinamide ribonucleotide transformylase with potent cell growth inhibition. *J. Med. Chem.* **1997**, *40*, 2502−2524.
(79) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; et al. Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]-indole-*N*-acetic acid (lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications. *J. Med. Chem.* **2005**, *48*, 3141−3152.
(80) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *48*, 3714−3728.
(81) Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K. Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture. *J. Am. Chem. Soc.* **2005**, *127*, 4632−4639.
(82) Wang, J.; Kang, X.; Kuntz, I. D.; Kollman, P. A. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *J. Med. Chem.* **2005**, *48*, 2432−2444.
(83) Tikhe, J. G.; Webber, S. E.; Hostomsky, Z.; Maegley, K. A.; Ekkers, A.; et al. Design, synthesis, and evaluation of 3,4-dihydro-2*H*-[1,4]-diazepino[6,7,1-*hi*]indol-1-ones as inhibitors of poly(ADP-ribose) polymerase. *J. Med. Chem.* **2004**, *47*, 5467−5481.
(84) Ealick, S.; Babu, Y.; Bugg, C.; Erion, M.; Guida, W.; et al. Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 11540−11544.

JM0608356