A Critical Assessment of Docking Programs and Scoring Functions

Gregory L. Warren,*,† C. Webster Andrews,‡ Anna-Maria Capelli,# Brian Clarke,|| Judith LaLonde,†,§ Millard H. Lambert,‡ Mika Lindvall,⊥,● Neysa Nevins,† Simon F. Semus,† Stefan Senger,⊥ Giovanna Tedesco,# Ian D. Wall,|| James M. Woolven,⊥ Catherine E. Peishoff,† and Martha S. Head†

GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina 27709, GlaxoSmithKline, Centre Via Alessandro, Fleming 4, 37135, Verona, Italy, GlaxoSmithKline, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, U.K., and GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K.

Received April 17, 2005

Docking is a computational technique that samples conformations of small molecules in protein binding sites; scoring functions are used to assess which of these conformations best complements the protein binding site. An evaluation of 10 docking programs and 37 scoring functions was conducted against eight proteins of seven protein types for three tasks: binding mode prediction, virtual screening for lead identification, and rank-ordering by affinity for lead optimization. All of the docking programs were able to generate ligand conformations similar to crystallographically determined protein/ligand complex structures for at least one of the targets. However, scoring functions were less successful at distinguishing the crystallographic conformation from the set of docked poses. Docking programs identified active compounds from a pharmaceutically relevant pool of decoy compounds; however, no single program performed well for all of the targets. For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity.

Introduction

In the past decades the number of protein structures publicly available in the Research Collaboratory for Structural Biology (RCSB) database has grown from one structure in 1972 to approximately 30 000 protein structures currently, with thousands being added each year.[1] This figure does not include the large number of proprietary structures held by pharmaceutical and biotechnology companies. Hand in hand with this growth in available protein structural data has come an increase in the number of compounds available in a form appropriate for virtual screening, both actual small molecules available in corporate and public compound collections and virtual small molecules accessible through computational enumeration of virtual library templates. An open question remains of how to best make use of these data and how to obtain the maximum value from these structural and synthetic investments using computational methods that are both theoretically grounded and pragmatically useful.

The aim of the study described here was to survey the current state of technology for structure-based drug design, focusing specifically on docking and scoring algorithms. A growing number of evaluations of docking programs and scoring functions have been published in recent years, including validations of new methods, head-to-head comparisons of docking programs,[2−18] and studies examining correlations between docking scores and compound affinity.[19−26] This study differed from these primarily in two ways. First, we examined the performance of many docking programs across a range of target types. Second, the compound set for each target was made up of a large number of closely related compounds for which experimental affinities have been measured using a standard protocol, generally by a single research group. This study measured the performance of docking and scoring algorithms on three tasks of particular relevance to drug discovery: prediction of protein-bound conformations, virtual screening for lead identification, and potency prediction for lead optimization.

For the evaluation described here, we compared as many docking programs as possible, including software currently licensed at any GlaxoSmithKline Pharmaceuticals (GSK) research site and supplemented by software for which vendors were willing to provide temporary licenses. The docking evaluation was carried out for eight protein targets of interest to GSK. For each protein target, we collected a high-quality data set containing only pharmaceutically relevant small molecules. All compounds in the data sets were synthesized in support of active GSK targets, and all compound classes represented in the data sets have shown biological activity in in vitro assays. We did not include decoy compounds selected from public databases such as the ACD or WDI. By selecting compounds in this way, we built a combined data set that closely mimics typical corporate compound collections. For each protein target/compound set, we had a number of crystallographically determined protein/ligand structures, ranging from 6 for 1 target to a maximum of 54 for another. Furthermore, the evaluation was set up to give programs the best possible opportunity of performing across a diverse, carefully compiled data set. For all protein targets, a GSK computational chemist experienced with a specific protein target provided up-front guidance concerning details of the protein binding site. For each docking algorithm, a GSK computational chemist with expertise with a specific program used that program optimally. By organizing the evaluation in this manner, this study characterized the state

* To whom correspondence should be addressed. Phone: (610) 917-5153. Fax: (610) 917-4206. E-mail: Gregory.L.Warren@gsk.com.
† Collegeville, PA.
‡ Research Triangle Park, NC.
# Verona, Italy.
|| Essex, U.K.
§ Current address: Locus Pharmaceuticals Inc., Four Valley Square, 512 Township Line Road, Blue Bell, PA 19422.
⊥ Hertfordshire, U.K.
● Current address: Chiron Corporation, 4560 Horton Street, Mailstop 4.2, Emeryville, CA 94608.

10.1021/jm050362n CCC: $33.50
Published on Web 08/13/2005

Table 1. Protein and Ligand Data Set Details

protein	target type	no. of ligands	no. of ligand classes	no. of cocrystals	max affinity (nM)	min affinity (nM)
Chk1	kinase	193	2	15	7	>10000
factorxa	serine protease	218	4	10	<1	5000
gyrase B	isomerase	138	3	7	4	>10000
HCV polymerase	polymerase	205	2	13	5.6	>10000
Met tRNA synthetase	synthetase	144	2	31	1	>10000
E. coli PDF	metalloprotease	199	3	2	1	>10000
Strep PDF	metalloprotease	186	3	4	<2	>10000
PPARδ	nuclear hormone receptor	206	5	54	0.3	>10000

of the art for a wide range of docking algorithms and scoring functions applied to systems of relevance for drug discovery. The results of this study provided us with a benchmark against which we can measure future progress along with a validated data set that can be used to evaluate a wide range of computational technologies beyond docking and scoring.

The organization of this docking algorithm evaluation paper is as follows. The next section describes the evaluation process, including details of the protein and ligand sets, along with a general overview of how the docking algorithms were applied to the data sets. The third section contains detailed results of the evaluation, with discussion in the fourth section. Specific computational details and parameters for each program are included in the Methods section at the end of the manuscript, while detailed tabulated results and graphs for every docking program/protein target pair are included in Supporting Information.

Evaluation Methodology

Protein Targets. The docking evaluation was carried out for eight specific proteins (the "targets") of seven protein types (Table 1). The proteins were chosen to include targets of active interest within the pharmaceutical industry and to encompass a variety of modes of action, binding site shapes, and chemical characteristics.

Chk1 kinase is a serine/threonine protein kinase of the CAMK family[27] and is responsible for cell-cycle arrest in response to DNA damage.[28] Inhibition of Chk1 kinase is therefore an attractive target for enhancing the action of DNA-damaging cytotoxic agents in the treatment of cancer. Multiple in-house crystal structures are available for CHK1/inhibitor complexes. In all cases, the inhibitors bind in the ATP binding pocket and form key hydrogen bond interactions with the protein backbone.

Factor Xa is a trypsin-like serine protease and is a key enzyme in the coagulation cascade.[29] Factor Xa initiates fibrin clot formation through the activation of prothrombin to thrombin and is a target for the treatment of thrombosis. Much of the substrate binding site is a shallow solvent-exposed groove, with the exception of the deep S1 pocket which preferentially binds the positively charged lysine and arginine amino acids. Inhibitors of factor Xa bind in this shallow substrate binding groove and span the S1−S4 subsites.

Gyrase is a bacterial type II topoisomerase involved in DNA replication, repair, recombination, and transcription.[30] The gyrase A2/B2 tetramer utilizes ATP hydrolysis to negatively supercoil DNA, with the ATPase activity located in the gyrase B subunit. The gyrase B inhibitors included in this evaluation all overlap portions of the ATP binding site and are competitive with ATP. However, these inhibitors are not ATP mimetics and bind quite differently from kinase inhibitors binding in kinase ATP sites.

Methionyl tRNA synthetase (MRS), an antibacterial target, is a class I amino acid tRNA synthetase and loads methionine amino acids onto tRNA for use in protein synthesis.[31] MRS is a homodimer; each monomer contains a Rossman fold domain typical of class I tRNA synthetases. Independent binding subsites for tRNA, ATP, and methionine are located in the MRS monomers. The methionine binding site has been delineated in a publicly available crystal structure.[32]

Hepatitis C RNA polymerase NS5B (HCVP) is an essential nonstructural enzyme responsible for replication of viral RNA.[33] There is no functional counterpart to this enzyme in the human genome, making HCVP an attractive target for treating hepatitis C infection. As with other polymerases, the protein fold and domain arrangement can be illustrated by the palm, thumb, and fingers of a right hand. HCVP differs from many other polymerases in that the ends of the thumb and finger domains are in contact. Because the HCVP binding site accommodates nucleotide, template, and product, the protein has a large binding site surface making HCVP a particularly challenging target for docking algorithms.

Polypeptide deformylase (PDF) is a bacterial enzyme that removes an N-terminal formyl group from newly synthesized proteins to generate mature, active proteins.[31] This deformylation reaction is a required step in bacterial protein synthesis but is not required in mammalian systems. PDF is therefore an attractive antibacterial target. PDF is a metalloprotease that carries out the same chemistry as matrix metalloproteases (MMP). In contrast to the MMPs, PDF is constricted in size near the metal binding site and does not have prime-side substrate binding pockets. A large number of high-resolution (<2.0 Å) public and in-house PDF structures from multiple bacterial species are available. For this docking evaluation, we have included PDF orthologues from *E. coli* and *S. pneumococcus*.

Peroxisome proliferator-activated receptor δ (PPARδ) is a nuclear hormone receptor that plays a role in lipid metabolism. Agonists of PPARδ have potential therapeutic value for the treatment of metabolic disorder.[34] PPARδ is activated in vivo by saturated and unsaturated long-chain fatty acids. The binding site for these amphipathic ligands is a largely hydrophobic cavity with a specific acid recognition element enclosed within the protein surface.

Protein Preparation. The protein structures used in this evaluation were selected by computational chemists with expertise in each particular protein target (the "system experts") and were prepared for docking calculations by a single computational chemist. Once prepared, the target structures were passed on to computational chemists with expertise in particular docking algorithms (the "program experts"). To avoid inadvertent bias in the calculations, there was as little overlap as possible between the group of system experts and the group of program experts.

For each protein target, the system expert selected a representative protein structure to be used for all docking calculations. The system expert therefore took special care to select a structure that both was a high-quality structure of good resolution and

also could accommodate all relevant compound classes. Binding site residues and amino acid ionization states were identified using automated methods, and the automated definitions were modified where necessary by the system expert. The system expert further provided guidance concerning any crystallographically identified waters considered important for compound binding. This collection of information was passed on to each of the program experts for use in setting up docking calculations. This process for preparing and distributing the docking structures was designed to achieve the best possible algorithmic performance while reducing the influence of known answers on the calculations.

Ligand Sets. Our aim was to generate a data set that closely represented a typical pharmaceutical compound collection. System experts therefore selected and compiled ligand sets for each protein target based on the following general guidelines:

(1) The set should include 150−200 compounds for each protein target.

(2) There should be two or three congeneric series per protein.

(3) A cocrystal must exist for at least one representative from each compound class.

(4) Ligand affinities for a given protein target should have been measured using a consistent assay format, and those affinities should span at least 4 orders of magnitude.

(5) Inactive compounds should make up less than 20% of the final set for each target. Extremely active compounds should similarly represent less than 20% of the target set.

Table 1 lists the characteristics of the compound sets used in this docking and scoring evaluation. Chemical structures are shown in Figure 1 for a representative of each compound class. The order in which compound classes appear in this figure corresponds to the order of protein targets in Table 1; e.g., compounds **1** and **2** in Figure 1 are representatives of the Chk1 kinase compound classes.

For the most part, the selection guidelines were met for all protein targets. The number of compounds per target ranged from 138 to 219, with two or three compound classes for six of the eight targets. The combined PDF compound set contains 199 compounds; affinities for 186 of these have been measured for both *E. coli* and *S. pneumococcus* PDF. In the case of factor Xa, the compound set was expanded to include a small number of compounds for which cocrystal structures are publicly available. In the case of PPARδ, the boundaries delineating distinct congeneric series are somewhat arbitrary. The number of compound classes is therefore larger in order to encompass compound class differences as fully as possible. In every instance but one, there was at least one cocrystal structure for each compound class, with multiple cocrystals for most classes. In *E. coli* PDF, one of the three classes did not have a cocrystal structure, but there were two cocrystals for that class in *S. pneumococcus* PDF. The requested affinity ranges were met for all of the target ligand sets, although not necessarily for each compound class within a given target set.

Ligand Preparation. Once compiled, the complete set of 1303 compounds was passed on to a single computational chemist who prepared the compounds for docking calculations. A single good starting geometry was generated from an input SMILES string. This starting conformation is not guaranteed to be a global minimum under any molecular mechanics force field but is guaranteed to have reasonable bond distances and angles and correct atom hybridization. From this starting point, four final SD files were prepared, with variations in the treatment of hydrogens and ionization: (1) all hydrogens, acids, and bases ionized for pH 7, (2) polar hydrogens only, acids and bases ionized, (3) all hydrogens, acids and bases nonionized, and (4) polar hydrogens only, acids and bases nonionized. These four ligand SD files were distributed to the program experts, who selected the small molecule representation most appropriate for a particular docking algorithm.

Table 2. Docking Protocols Included in This Evaluation

docking program	alternative protocols
Dock4	chemistry
	contact
	energy
DockIt	
FlexX	FlexX score
	DrugScore
Flo	Mcdock
	Mcdock+
	Fulldock
	Sdock
	Zdock
Fred	ChemScore
	ScreenScore
Glide	
Gold	
LigFit	CVFF
	Dreiding
MOE	
MVP	

Docking Algorithms. The docking and scoring evaluation described here aimed to include as many docking programs as possible used as expertly as possible. We evaluated all docking programs available under GSK's current licensing arrangements. The set of already licensed programs was augmented to include readily available docking programs for which the vendor was willing to provide a temporary demo license. The 10 programs evaluated are listed in Table 2. In addition, some of the programs evaluated offer a choice of multiple scoring functions or docking algorithms to drive the generation and selection of docked poses (as indicated in the second column of Table 2), resulting in a total of 19 docking protocols.

To optimize the performance of each docking program, computational chemists with expertise in a particular program were identified from the worldwide GSK computational chemistry community. Each program expert was given complete freedom and sufficient time to maximize the performance of the docking program. In all cases, but especially for those programs evaluated under a demo license, consultation with software vendors was encouraged. The vendors were not able to see any of the protein targets or ligand structures, but were able to provide guidance concerning computational details of the program itself. No time deadlines were imposed so that even low-throughput docking programs could be evaluated. Indeed, no constraints whatsoever were placed on the level of agonizing over details of how each docking protocol was applied.

Analysis Measures. The evaluation focused on three typical uses of docking programs: (1) prediction of conformations of small molecules bound to protein targets, (2) virtual screening of compound databases to identify leads for a protein target, and (3) prediction of compound affinities to guide lead optimization efforts. For each of these typical uses, we conducted separate docking calculations and analyzed the results using different analysis measures.

Prediction of Protein-Bound Conformations. Two measures were used to assess the similarity of all docked poses to the crystallographically identified bound orientations. As the primary analysis measure, a symmetry-corrected root-mean-squared deviation (rmsd) was computed for ligand heavy atoms. Results are reported here for compounds docked within 2 and

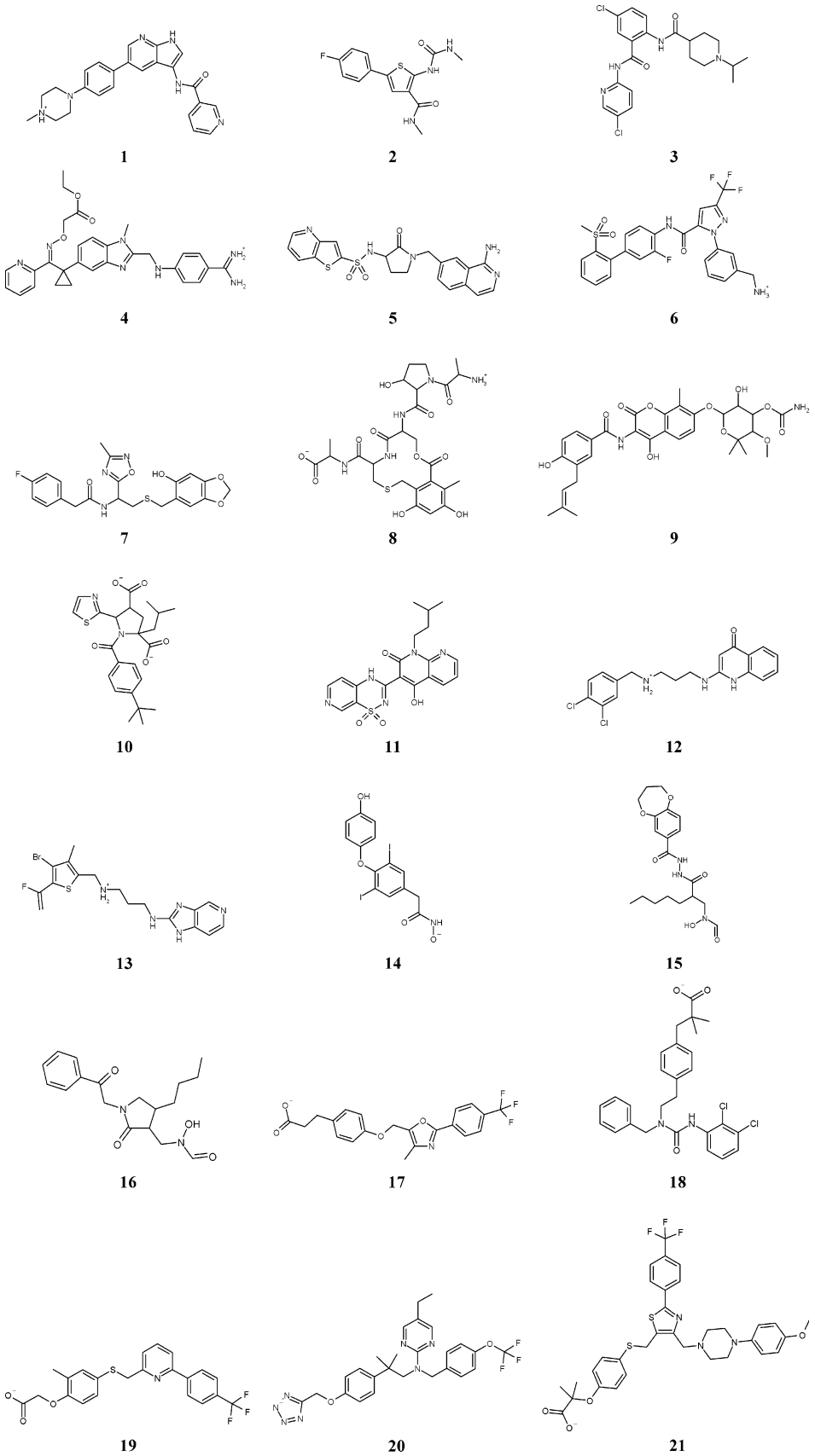

Figure 1. Representative molecules from the 21 compound classes in the ligand data set.

4 Å of the crystal conformation. These cutoff values were selected based on visual inspection of many docked poses. Within the 4 Å cutoff, docked poses were located within the binding site in a roughly correct global orientation. Within the 2 Å cutoff, docked poses were oriented properly within the binding site, and details of the binding were predicted well

enough to be useful for compound design. As a supplementary analysis tool, a volume overlap Tanimoto similarity index T_{vol} was computed as defined by

$$T_{vol} = \frac{O_{X,D}}{I_X + I_D + O_{X,D}} \quad (1)$$

where I_X and I_D are self-volume overlap integrals for the crystal and docked conformations and $O_{X,D}$ is the volume overlap between crystal and docked poses. If there is no overlap between a docked pose and the crystal conformation, $T_{vol} = 0$. Conversely, $T_{vol} = 1$ would indicate a perfect overlap of the two orientations. In practice, no docked poses attained this perfect overlap because initial molecule conformations were generated from SMILES strings. The bond angles and distances generated by this method were reasonable but will not perfectly match the values seen in the crystal structure because different molecular mechanics force fields were used. For the docked poses generated across the targets, T_{vol} ranges between 0.5 and 0.99 for ligands docked within 2 Å of the crystal orientation (data not shown). In conjunction with the computed rmsd, T_{vol} presented a fuller picture of the agreement between docked and crystal conformations. In particular, this measure allowed the identification of docked poses that agree with the crystal result in most instances but have small details incorrect. For example, a compound docked correctly except for the rotation by 180° of an asymmetrically substituted phenyl ring would lead to a relatively large rmsd value, while T_{vol} would correctly indicate a high level of similarity between the two orientations.

Docking as a Virtual Screening Tool. Two measures were used to assess the ability of docking algorithms and their scoring functions to identify active compounds from a pool of decoys for a particular target. The first of these measures determined how quickly active compounds were identified compared to random chance. This measure, designated enrichment, is the signal-to-noise ratio. In this case, success was declared if an algorithm was able to identify at least 50% of the active compounds within the top 10% of the score-ordered list, giving an enrichment above random of 5. This value of 5 represented an enrichment at least halfway between random and the theoretical maximum enrichment for the data sets used in this evaluation. The second measure of success, designated lead identification, is a measure of cost. This measure asked how many compounds must be screened before at least one active representative of every active compound class has been identified. For lead identification we do not need to find all the active compounds, only one or two representatives from each class. The molecular data set used in this evaluation contained two to five active congeneric series for each target. Our lead identification measure determined whether a docking algorithm preferentially identified or missed compound series. For lead identification, success was declared if all active compound classes were identified within the top 10% of the score-ordered list. In addition, boost plots of percent actives found versus percent compounds screened in the docking-score-ordered list for all targets were generated to aid in the comparison of docking algorithm performance within and across targets. Plots of initial enrichment rates along with the complete set of boost-plot data are included in Supporting Information.

Scoring as an Affinity Prediction Tool. Mathematical comparisons were made between experimentally measured compound affinities and calculated docking scores in order to assess affinity predictions for each target. Measured affinity was compiled for each compound as IC_{50}, EC_{50}, or K_i. For each target, the affinity measurements are of a single type: IC_{50}, EC_{50}, or K_i. There is no mixture of measurement types within a single target. We have converted the measured affinity to pA (=pAffinity), as defined in

$$pA = -\log\left(\frac{A}{1\ \mathrm{M}}\right) \quad (2)$$

where A is the measured IC_{50}, EC_{50}, or K_i and the reference state is 1 M. To allow direct comparisons of scores from different docking programs, autoscaled docking scores S_i' were computed as defined in

$$S_i' = \frac{-|S_i| - (-|\bar{S}|)}{\sigma} \quad (3)$$

where S_i is the score generated by a single docking program/scoring function pair for a single docked compound, $\bar{S}$ is the average score for a single docking program/scoring function pair applied to all compounds for a given target, and σ is the standard deviation of these docking scores. For each target in the evaluation, we have generated graphs of pAffinity versus scaled docking score for all of the docking programs and scoring functions. Because measured affinities and calculated scores have been transformed as defined in eqs 2 and 3, these graphs can be compared directly to visually assess the ability of scoring functions to predict compound affinity. To mathematically assess predictions of compound affinity, we have computed a linear correlation coefficient r as defined in

$$r = \frac{\sum_i (S_i' - \bar{S}')(pA_i - \overline{pA})}{\sqrt{\sum_i (S_i' - \bar{S}')^2}\sqrt{(pA_i - \overline{pA})^2}} \quad (4)$$

where S' is the scaled docking score. For comparisons between pAffinity and scaled docking score, a correlation coefficient $r = -1$ would correspond to a perfect correlation between compound affinity and docking score while $r = +1$ would mean that the scoring function was universally ranking less active compounds higher than more active compounds.

Experimental Design and Aims of the Evaluation. As described above, the program experts were provided with a package of information, including a careful delineation of the residues encompassing the binding site as well as commentary from the system expert concerning subtleties of the target structure. However, program experts were not provided with any example cocrystal structures; indeed, it was strongly requested that no one seek out such structural data. The program experts were nevertheless free to make use of their own generic understanding of the target types; e.g., incorporating algorithmic procedures for docking into the multiple subsites of a serine protease or for correctly orienting metal binding groups in a metalloprotease. This experimental design sought to reproduce a specific real-life situation: We have a protein structure and perhaps a lead molecule from high-throughput screening. Can we predict a priori how that lead sits in the protein binding site in order to drive early optimization efforts? In the absence of a lead molecule, can we identify potential leads through docking-based virtual screening? Can we use docking and scoring to rank-order compounds during lead optimization? In addition, this experimental design avoided inadvertent driving of the calculations toward a known answer, thereby leveling the

Table 3. Docking Protocol That Produced the Greatest Number of Correctly Docked Structures

	Dock4	FlexX	Flo	Fred	LigFit
kinase	energy	DrugScore	Mcdock	ScreenScore	ChemScore
protease	energy	Flexx	Fulldock	ChemScore	
isomerase	energy	DrugScore	Mcdock	ChemScore[a]/ScreenScore[a]	ChemScore
polymerase	energy		Fulldock[b]	ScreenScore[c]	ChemScore
synthetase	energy	Flexx	Fulldock	ChemScore	
metalloprotease	energy	Flexx	Mcdock	ChemScore[d]	Dreiding
NHR	contact	DrugScore	Mcdock+	ScreenScore	ChemScore

[a] Docked using hypothetical reconstructed loop structure. [b] One water included in binding site definition. [c] Two waters included in binding site definition. [d] Pharmacophore matching in metal-chelating region.

playing field as much as possible by allowing each docking program to compete on the basis of the same core set of information.

Results and Discussion

This evaluation examined three specific uses of docking programs; the results for each use are presented separately here. In results section A we assess the ability of docking programs to generate and identify crystallographically determined bound orientations of compounds for which we have protein/ligand crystal structures. In section B we assess the ability of docking programs to identify active compounds from a decoy pool and further examine whether we are able to enrich hit rates for the right reasons. In section C we examine the most difficult challenge for docking programs, prediction of binding affinity for a large number of closely related compounds. In all sections we make specific observations followed by evaluation results that support those observations.

A. Prediction of Protein-Bound Conformations. A.1. Docking programs were able to generate crystal conformations. Nineteen docking protocols were used to predict bound conformations for the 136 compounds for which we have protein/ligand crystal structures. Each docking protocol returned multiple docking poses for each ligand; rmsd values were computed for all poses returned. Statistics were compiled for the best rmsd for a given compound/docking protocol pair, without consideration of where that pose was ranked in the list of all poses returned by the docking protocol. Because of the small number of PDF structures included in the data set, results for *E. coli* and *S. pneumococcus* PDF were combined. The best rmsd results for the target types are reported in the top panels of each part of Figure 2. For each program, black bars denote the percentage of ligands for which any docked pose was within 2 Å of the crystal conformation while gray bars indicate the additional percentage of ligands docked within 4 Å. Where multiple protocols were possible for a given docking program, the single best result is included in the figures; the selected docking protocols are indicated in Table 3. Complete results are tabulated in Supporting Information.

As shown in Figure 2, overall success rates were quite good across the protein targets. For all targets except HCVP, at least one program was able to dock ≥40% of the ligands within 2 Å of the crystal conformation. For five of the seven targets, at least one program docked ≥50% of the ligands well. Indeed, for several protein targets, 90% of the ligands could be docked in the correct orientation and 100% could be docked in the correct location. Clearly, docking algorithms were able to explore conformational space sufficiently well to generate correctly docked poses.

For targets Chk1 kinase and PDF, good performance was seen across many docking protocols, with six docking protocols able to dock ≥50% of the compounds within 2 Å of the crystallographic conformation. In both cases, compounds are bound in relatively small, well-defined binding sites and make a small number of key orienting interactions with protein atoms. As the size of the binding site or complexity of the compounds increased, fewer protocols were able to generate poses close to the crystal conformation. In particular, the binding site of HCVP is extremely large, encompassing NTP, template, and product binding regions. The system expert focused the search space to the NTP subsite, but even this guidance left a large amount of protein surface in the search space, and no docking program was able to generate docked poses within 2 Å for ≥40% of the compounds. Conversely, in the case of PPARδ, the binding site is not particularly large. In this instance, the hydrophobicity and conformational complexity of the compound classes (**17**−**21** in Figure 1) may have affected the ability of docking programs to identify good poses. Figure 3 plots rmsd versus T_{vol} for the best-scoring poses returned by Gold. For a substantial population of docked conformations, rmsd values of 6−10 Å were seen for poses that overlap significantly ($T_{vol} \geq 0.5$) with the crystallographic conformation. The compounds were placed in the binding site but did not adopt small-molecule conformations that allow the compounds to be oriented correctly within the site.

A.2. Scoring functions were less successful at identifying the pose closest to the crystal conformation. Docking accuracy statistics for the top-scoring pose returned by all docking programs are shown in the bottom panels of each part of Figure 2. For all targets, when considering only the best-scoring pose for each compound, a smaller percentage of compounds were docked within 2 or 4 Å of the crystallographic conformation. In addition, although several of the docking programs reported multiple docking scores, none of these scoring functions were able to reliably identify the best-docked pose (data not shown). Although docking accuracy decreased for the top-scoring pose returned, this performance decrease was not as extreme as one might have expected a priori. For five of the seven target types, at least one docking program/scoring function pair was able to identify poses within 2 Å of the crystallographic conformation for ≥40% of the compounds.

A.3. Docking into a single protein structure did not create large difficulties for multiple compound classes. All of the docking calculations for this evaluation were carried out using a single crystal structure for each protein target. Although each system expert selected a structure that should accommodate all compound classes, docking a compound into a noncognate protein structure may have adversely affected chances for identifying correctly docked poses.

In Figure 4, rmsd values of ≤4 Å are plotted for compounds belonging to each compound class. Each column of the graph contains rmsd values for all poses generated for all compounds within a class. Vertical gray lines separate the compound classes belonging to each target type. Results for all docking programs are included in this graph.

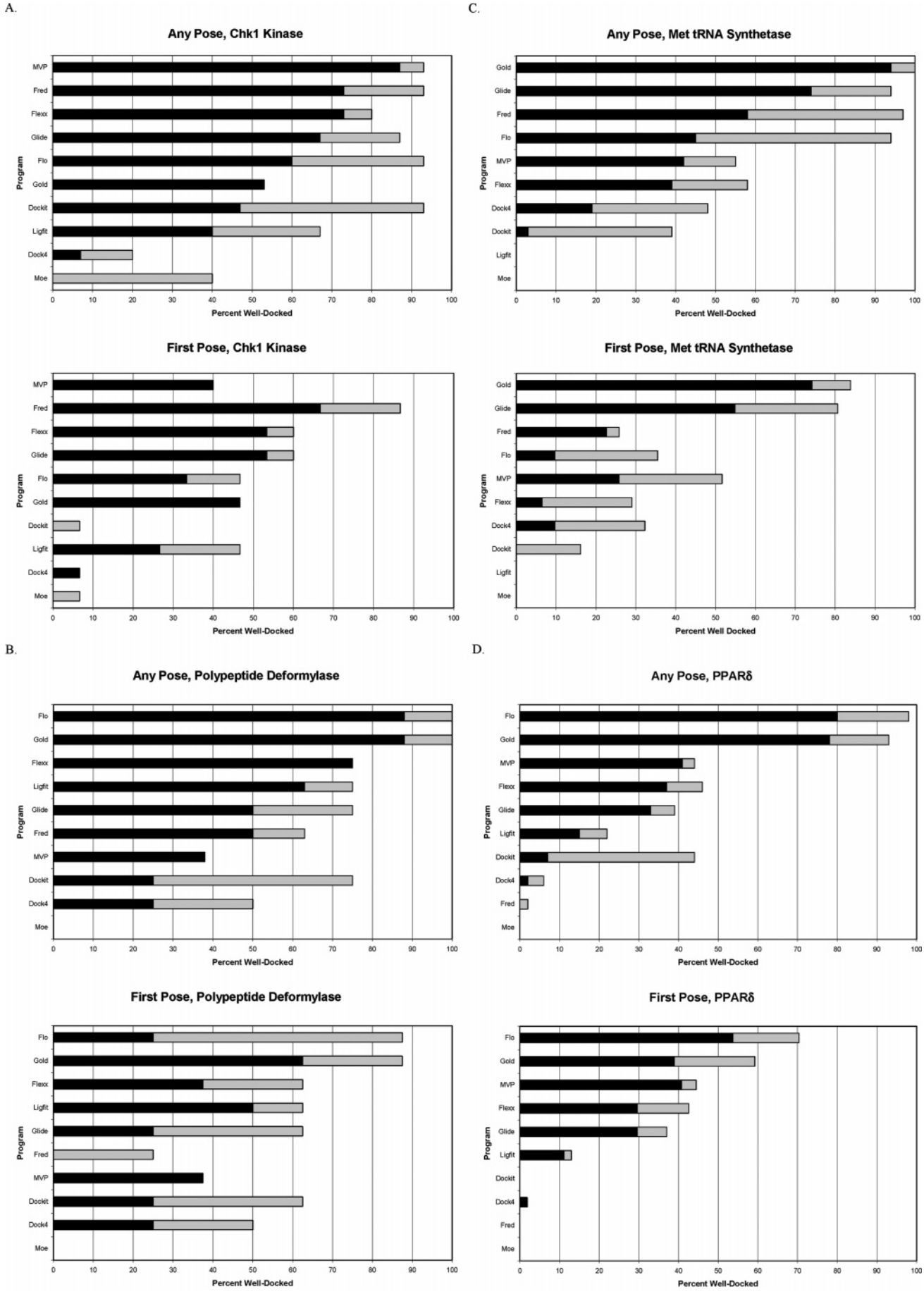

Figure 2 (Continued on next page)

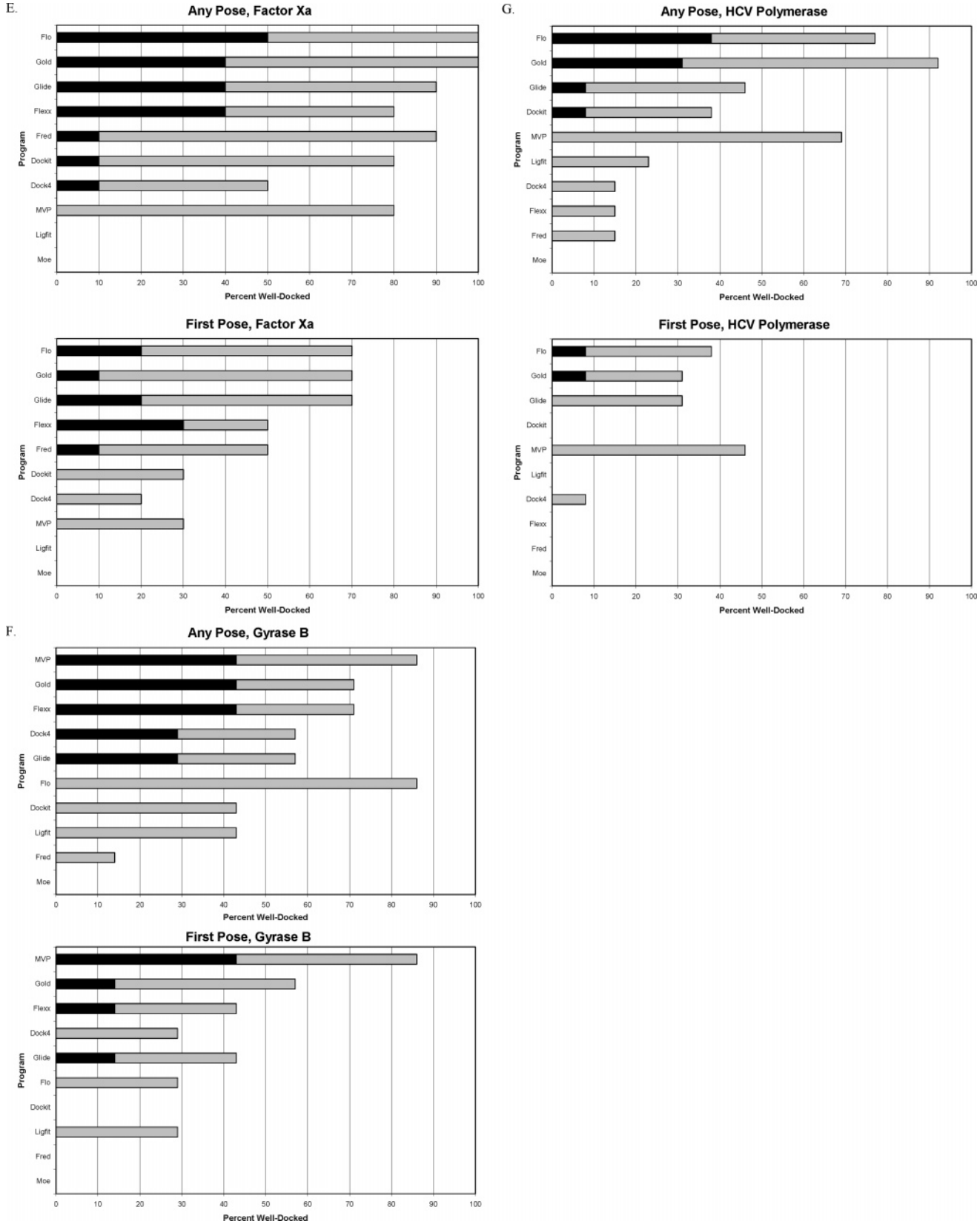

Figure 2. The rmsd results for all protein targets. Black bars indicate the percentage of compounds for which a docked pose was found within 2 Å of the crystal structure, while gray bars indicate the percentage of compounds for which a docked pose was found within 4 Å of the crystal structure. The first figure in each target section graphs the best rmsd for any pose returned by a particular docking program, and the second figure in each target section graphs the rmsd for the first pose returned by a particular docking program. Docking programs were able to reproduce experimentally determined protein-bound conformations in that at least one docking program placed ≥50 first poses within 2 Å for four of the seven targets evaluated. However, performance by any docking program was not consistent as noted by the fact that the program with the best performance, listed first on each of the protein target graphs, changes.

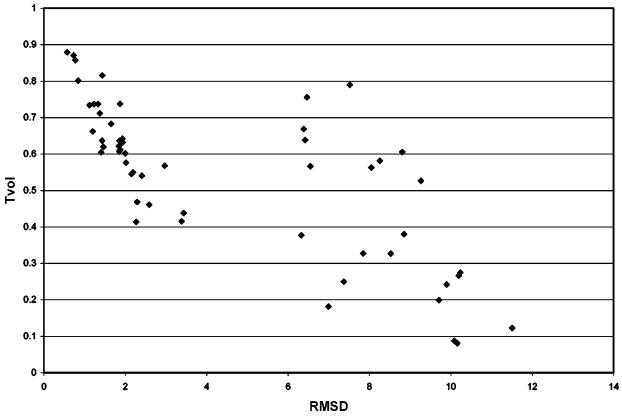

Figure 3. Comparison of rmsd versus T_{vol} for PPARδ for the first pose returned by the docking program Gold. For the poses from 2 to 4 Å there is a strong correlation between the rmsd and T_{vol}. However, a significant population of docked conformations between 6 and 10 Å have a high T_{vol} value ($\geq$0.5), indicating the compounds were placed in the binding site correctly but do not adopt the correct small-molecule conformation.

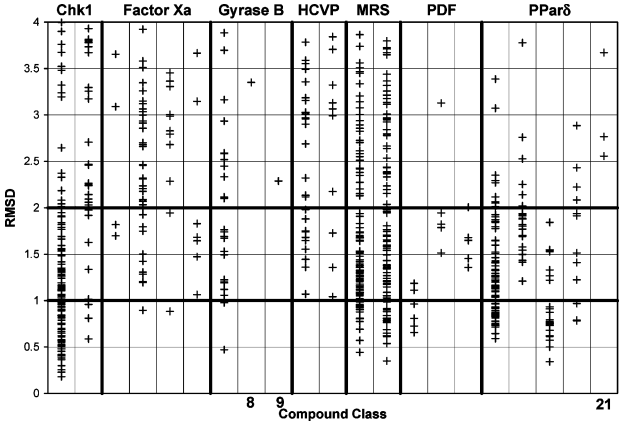

Figure 4. Plot of all rmsd values of $\leq$4 Å from all docking programs for molecules in each compound class. With the exception of classes **8**, **9**, and **21** all compound classes placed, by at least one docking program, more than one representative ligand within 2 Å of the crystal structure. For the protein targets in this evaluation, cross-docking multiple compound classes into a single-crystal structure was done successfully.

The graph in Figure 4 identifies only three compound classes for which no docking program could find a docked pose within 2 Å of the crystallographic conformation: gyrase B compound classes **8** and **9** and PPARδ compound class **21** (Figure 1). In all three cases, these compound classes contain features that would be expected to be particularly challenging.

Compound class **8** contains a heterocyclic macrocycle at the core of the molecule. Given that the starting conformation of this ring was randomized during ligand preparation and given the conformational complexity of such a macrocycle, it is unsurprising that no docking program was able to recover the crystal-bound conformation of the central ring (Figure 5). Nevertheless, despite the small-molecule conformational search issues, the docking program was able to correctly place the aromatic ring and aminopropanoate substituent. Although the rmsd for this pose is 3.3 Å, binding features were captured well enough that chemical insights could be derived from the results of a docking calculation. Similarly, the best docking pose for compound class **9** oriented the compound correctly within the binding pocket and captured important binding features. In this instance, many of the docking programs selected an extended conformation for the butenylbenzamide substituent. The conformational flexibility of PPARδ compound class **21** also led to small-molecule conformational search issues that made docking challenging for this class. Even so, the acid group was correctly located near the acid-group recognition site in the PPARδ binding pocket, and the more hydrophobic portions of the molecule were docked into the correct subsites.

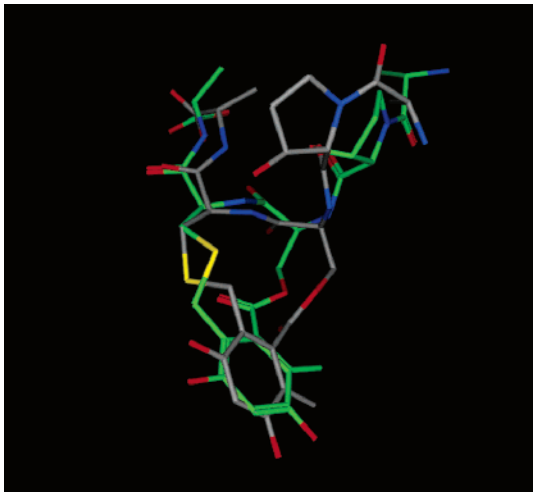

Figure 5. Comparison of the cocrystal structure determined protein-bound small-molecule conformation in gyrase B (shown using green carbon atoms) with the lowest rmsd pose (3.3 Å) generated by a docking program (shown using gray carbon atoms). While the conformation of the heterocyclic macrocycle is incorrect, the important pharmacophore elements, the aromatic ring and aminopropanoate, are placed correctly.

A.4. No single docking program performed well across all protein targets. In the top panels of each part of Figure 2, the results have been ordered such that the better-performing programs are located toward the top of the graph. A quick scan of the graphs reveals that different programs docked ligands well for all targets; there is no one program that is universally located near the top of the list.

B. Docking as a Virtual Screening Tool. The objective of this section of the evaluation was to determine how capable docking algorithms and their associated scoring functions are at selecting molecules active for a particular target from a pool of decoy molecules. We have shown in the previous section that docking algorithms could in many cases solve the search problem, i.e., could find the correct small molecule conformation and position the small molecule correctly within the protein binding site. At a coarse level, virtual screening is a test of the ability of scoring functions to differentiate between active and inactive chemotypes within the context of a protein binding site. In section C, we will examine at a finer level the ability of scoring functions to differentiate between active and inactive compounds of a single chemotype.

We would like to remind the reader that this study strove to evaluate docking algorithms under conditions similar to those used daily by a computational chemist for lead discovery. A conscious effort was therefore made to use a molecular test set that mimics that of a typical corporate collection, e.g., a large number of diverse chemical classes each of which contains a number active and inactive close chemical analogues. Compound sets containing congeneric members are typical both in corporate collections and in purchasable collections, real or virtual. The case examined here (where the decoy compounds include both diverse chemical classes and inactive close analogues) was inherently challenging. Thus, the results from such a study provided a measure of the state of the art under the most challenging and realistic of circumstances.

Table 4. Enrichment Factor for Actives (≤1 μM) Found at 10% of the Docking-Score-Ordered List

program	Chk1	FXa	gyrase B	HCVP	MRS	*E. coli* PDF	*Strep* PDF	PPARδ
ideal	10.0	9.8	10.0	9.5	10.0	7.6	8.3	8.6
Dock4	1.4	4.1	1.7	1.8	4.2	0.9	0.8	1.7
DockIt	4.2	2.0	2.0	1.0	1.0	0.2	0.0	3.2
FlexX	7.0	2.2	5.8	0.9	3.9	0.8	0.8	5.2
Flo+	5.6	2.7	2.3	3.4	1.7	1.5	0.8	3.6
Fred	2.9	4.1	1.9	2.0	0.6	3.2	1.2	1.1
Glide	6.3	3.4	1.0	1.0	5.3	0.6	0.4	4.8
Gold	0.1	4.1	4.0	0.0	0.8	1.0	0.1	5.5
LigandFit	3.3	1.9	2.8	1.8	2.9	2.9	1.7	1.2
MOEDock	3.9	0.6	0.0	0.0	1.0	2.1	0.6	0.0
MVP	7.2	5.8	5.3	3.6	6.4	6.7	6.9	3.9

Table 5. Percent of the Docking-Score-Ordered List Screened To Find at Least One Active (≤1 μM) Representative for All Compound Classes

program	Chk1	FXa	gyrase B	HCVP	MRS	*E. coli* PDF	*Strep* PDF	PPARδ
Dock4	44.5	92.0	45.2	1.4	0.2	6.3	1.8	12.7
DockIt	62.8	66.4	4.0	2.5	5.6	7.1	20.0	0.9
FlexX	7.6	51.6	49.6	2.0	0.5	10.4	10.6	4.7
Flo+	7.1	18.6	1.6	0.6	3.8	3.4	12.4	2.4
Fred	10.4	63.7	99.8	70.1	3.5	2.3	7.9	59.9
Glide	2.5	89.0	100	1.8	0.2	16.1	11.4	2.9
Gold	36.8	34.8	98.9	22.8	8.9	8.2	20.1	6.1
LigandFit	37.5	72.5	84.0	0.5	5.6	27.0	7.6	64.7
MOEDock	0.4	73.4	99.5	8.9	5.0	0.8	2.4	94.6
MVP	8.4	63.8	37.4	1.7	0.5	2.6	1.2	95.2

This data set differed from data sets used previously [4,6−8,11,18] in that it contained a high percentage of active compounds, from 6% to 13%. While this percentage of actives is not typical of molecule sets routinely used for virtual screening, this richness could be used to ask more detailed questions about performance. In particular, are docking algorithms quickly identifying all active chemotypes? Do inactive analogues confuse the algorithms and cause a decrease in performance? Are enrichment rates higher when the definition of active is skewed toward more potent molecules (≤100 nM) compared to the rates when less potent (≤10 μM) molecules are included?

B.1. Docking programs could identify molecules active against a target out of a population of decoy molecules. For all but the HCV polymerase target, at least one docking program/scoring function pair had an enrichment factor of 5 or greater (Table 4). For these seven proteins, performance generally fell into three broad categories; enrichment close to the theoretical maximum, intermediate enrichment values, and no or even negative enrichment (Figure 6A−G). The one exception was HCV polymerase where the best performance was in the intermediate range (Figure 6H). In this case, two programs had enrichment factors of roughly 3.5.

B.2. Docking programs could correctly identify all active chemotypes from a population of decoy molecules. Enrichment is a measure of performance that asks how quickly active compounds are found. However, it is not a measure of diversity or completeness. While finding active leads rapidly is important in the practical application of these algorithms toward virtual screening, an equally important measure of algorithm robustness is the ability to identify chemically diverse leads across diverse targets. Except for the serine protease factor Xa, at least one algorithm identified at least one member of all the active chemotypes within the top 10% of the docking-score-ordered list (Table 5). One program, Flo+, was able to identify at least one member of all active series within the top 20% of the docking-score-ordered list on all protein targets evaluated. With the exception of factor Xa, the success and consistency rates using the lead identification measure were considerably higher than performance based on enrichment alone.

Table 6. Ratio of Percent Screened of Score-Ordered List To Find at Least One Active Representative versus One Representative, Active or Inactive, for All Compound Classes

program	Chk1	FXa	gyrase B	HCVP	MRS	*E. coli* PDF	*Strep* PDF	PPARδ
Dock4	2.4	1.0	1.0	1.0	1.0	1.0	1.0	3.2
DockIt	1.2	1.0	1.0	1.3	1.0	1.0	1.0	1.0
FlexX	3.4	1.0	1.0	1.0	1.8	1.0	1.1	1.0
Flo+	5.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fred	1.0	1.0	1.0	1.0	2.1	1.2	1.0	1.3
Glide	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Gold	1.3	1.0	1.0	1.4	1.4	1.0	1.0	1.0
LigandFit	2.1	1.0	1.0	1.7	1.0	6.8	1.0	4.3
MOEDock	1.0	1.0	1.0	1.6	1.8	1.0	1.0	1.0
MVP	3.3	1.0	1.0	1.0	1.0	1.0	1.0	1.0

B.3. Inactive close analogues generally did not degrade lead identification performance. When a corporate compound collection is virtually screened, it is often difficult to differentiate between closely related active and inactive analogues. Because the virtual screening molecule set used by this evaluation contained inactive analogues, we were able to quantify the effect inactive analogues had on lead identification performance.

In Table 5 we list the percentage of the docking-score-ordered list that must be screened to find at least one active member of all active chemical classes for a particular target. This measure is designated the "percent to find actives". In addition, we examined the percentage of the docking-score-ordered list that must be screened to find at least one representative, active or inactive, for all active chemical classes. This measure is designated "percent to find classes". If a docking score misidentifies an inactive analogue and places it before active compounds in the score-ordered list, then the percent-to-find-classes number will be less than the percent-to-find-actives number. To better clarify where performance degradation occurred, we computed a ratio by dividing the percent-to-find-actives number by the percent-to-find-classes number. Where this ratio is greater than 1 (see Table 6) the docking score misidentified inactive compounds, resulting in a reduction in lead identification performance.

Of the eight protein targets evaluated, the largest reduction in lead identification occurred for Chk1 kinase. For Chk1, 7 out of 10 programs ranked inactive analogues above actives (Table 6). This result suggests that scoring functions were correctly identifying key interactions of kinase inhibitors with the ATP binding site but were not capturing more subtle compound differences that affect affinity. For all the targets, including Chk1, the algorithm that most rapidly identified at least one member of all active chemotype (Table 5) was not fooled by inactive analogues. The docking algorithm Glide, while not always identifying all active series rapidly (Table 5), had no inactive-analogue-induced reduction in lead identification performance.

B.4. With one exception, enrichment rates at 10% screened did not change when the definition of an active changed. For the results presented in Figure 6 and Table 4, actives were defined as compounds with better than micromolar activity. A further analysis of our virtual screening data set was carried out for all algorithms across all targets using different definitions of active, activity of ≤10 μM and activity of ≤100 nM. No significant changes in performance were observed as measured by enrichment (data not shown).

The one exception found was for the Chk1 kinase target. In this case, when active was defined as activity better than 100 nM, several programs (Dock4, DockIt, Gold, LigandFit) showed

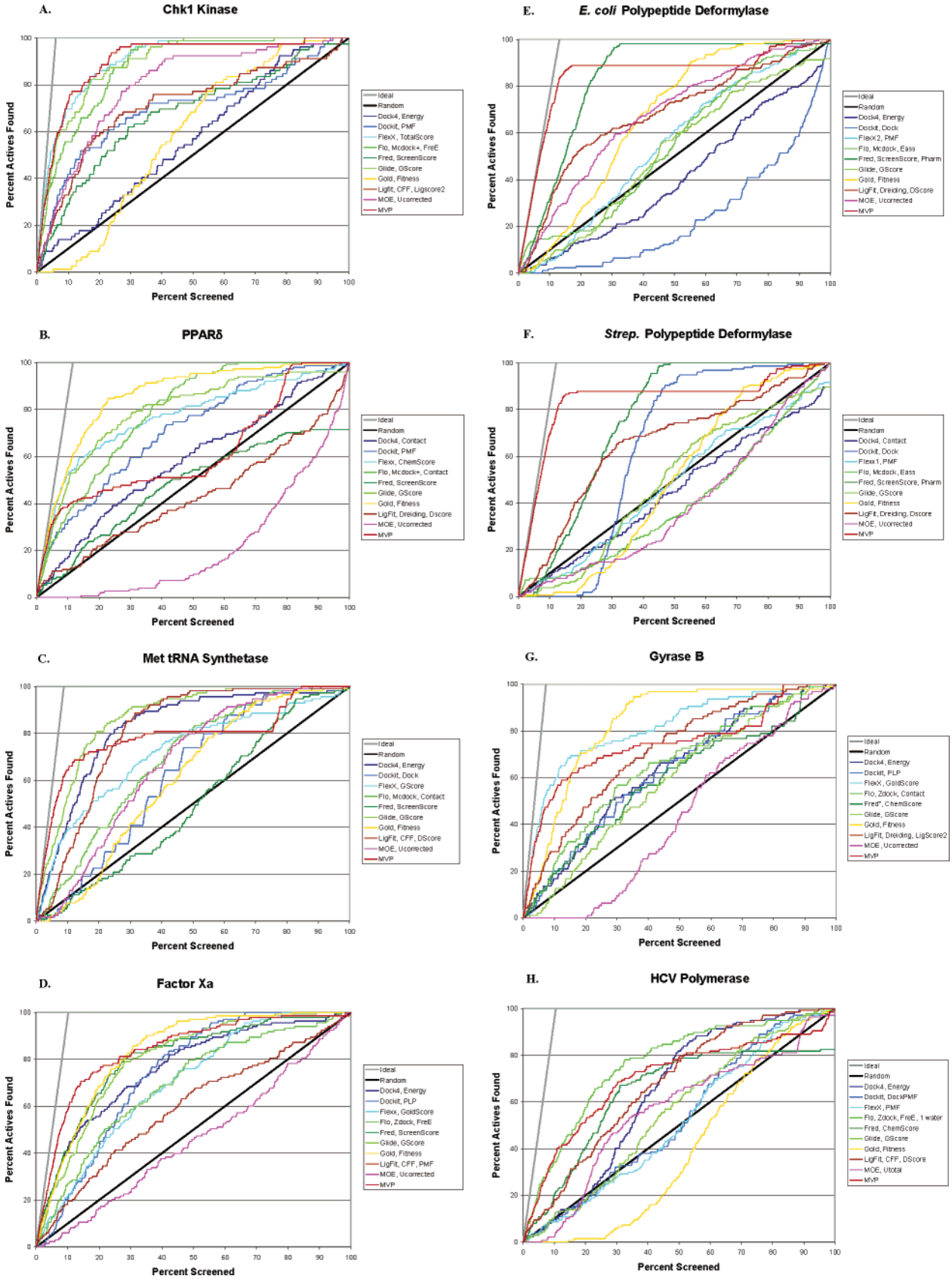

Figure 6. Boost plots of the percent active found versus percent of the docking-score-ordered list screened, using the scoring function with the highest enrichment at 10% screened for programs across all targets. The heavy black line represents the values expected if actives were selected at random. The heavy gray line represents the values expected if all active are placed sequentially at the top of the list. It is of interest to note that algorithm performance or enrichment varies dramatically across the targets.

a significant improvement in enrichment. This improvement appears to be due to differing performance of these docking algorithms for compound classes **1** and **2**. Of the two Chk1 kinase compound classes, only class **1** contained compounds

with better than 100 nM activity; no class **2** compounds showed activity of ≤100 nM. Using the ≤100 nM definition of activity, enrichment rates therefore measured the performance of these algorithms for class **1** only. When using the ≤1 μM activity definition, these algorithms were less successful at identifying compound class **2** as active for this target; more than 30% of the compounds had to be screened to find at least one active example of both chemical classes (Table 5). For these programs, the resulting enrichment values calculated using the more stringent activity definition that excluded the missed compound class **2** showed a dramatic improvement.

B.5. Enrichment and lead identification alone were not comprehensive criteria for determining algorithm performance. This evaluation has stressed measuring the performance of docking algorithms under standard working conditions. However, the criteria by which practical application performance is judged are and should be different from the criteria used for algorithm development. The ideal docking program, when applied to virtual screening, should be able to quickly identify all active compounds and all chemotypes across a diverse set of target proteins. Neither the enrichment nor lead identification criteria alone are adequate for determining ideal performance. For five of the targets evaluated in this study, at least one algorithm met both the enrichment and lead identification criteria for success as set by this evaluation. An exception was factor Xa where four docking algorithms had an enrichment factor greater than 4 at 10% screened yet required screening of more than 30% of the compound list before identifying at least one member of all the active chemotypes. Again, for gyrase B, three algorithms had enrichment at or greater than 4 and again required screening of more than 30% of the compound list before identifying active members of all chemotypes. A last example comes from the HCV polymerase target data. No docking algorithm met the enrichment criteria for success against this target. However, with only two exceptions all the algorithms successfully met the lead identification criteria (Table 5). Thus, we have demonstrated that neither enrichment or lead identification alone are sufficient measures for optimizing algorithm performance during development.

B.6. While docking programs could correctly identify active molecules, performance across diverse targets was inconsistent. We have shown that at least one docking algorithm could correctly identify a majority of the active compounds in the top 10% of the docking-score-ordered list of compounds for seven of the eight targets examined. In addition, we have shown that at least one program for each target could correctly identify all the chemical templates active against the target from a pool of decoy compounds. However, there was no single program that met either criteria for success for every target. In fact, there are many examples where a program had an enrichment factor greater than 5 at 10% screened for one target and an enrichment of 1 or less for another target. While there were examples of this for all of the algorithms evaluated, we illustrate the point using data from the program Glide (Figure 7 and Table 4) where for half of the protein targets evaluated Glide had enrichment rates of ≥3.4 at 10% screened and for the other half Glide had enrichment rates at or less than random.

The inconsistency in enrichment performance was mirrored in our lead identification performance measure (Table 5). No program was able to place at least one active molecule in the top 10% of the docking-score-ordered list for all 21 chemotypes. For many algorithms, performance fluctuated dramatically across the protein targets studied. The inconsistency in performance observed in this evaluation suggests that in the absence of training or validation data it is impossible to tell a priori whether a docking program will be successful at virtual screening for a given target.

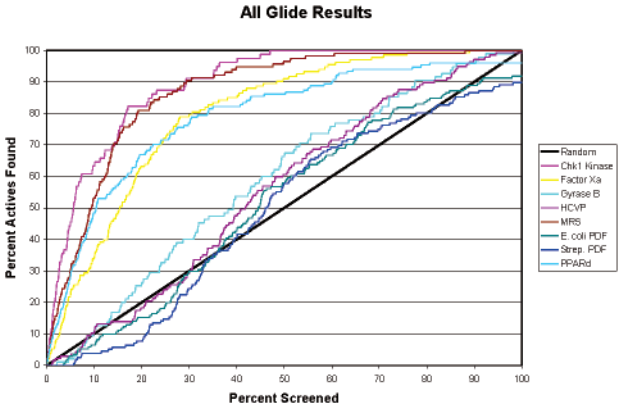

Figure 7. Illustrative example of how enrichment by a single program varied across the targets evaluated using data from the program Glide. Similar variation in performance was observed in all docking programs evaluated.

B.7. Good performance in reproduction of experimentally determined binding modes did not impart success in virtual screening. For five of the seven targets evaluated, more than 70% of the top-ranked poses are within 4 Å of the experimentally observed binding mode, and for four of the targets greater than 55% of the top-ranked poses were within 2 Å of the crystallographic conformation. However, we did not see a consistent correlation between the ability to reproduce binding modes and enrichment in virtual screening for lead identification. In these evaluation results, we saw examples of each of the four possible outcomes: (1) Cognate compounds were well-docked, and enrichment rates were high, e.g., Chk1/FlexX with an enrichment rate of 7.0. (2) Cognate compounds were well-docked, but enrichment rates were low, e.g., *E. coli* PDF/Gold with an enrichment rate of 1.0. (3) Cognate compounds were poorly docked, and enrichment rates were low, e.g., HCVP/FlexX with an enrichment rate of 0.9. (4) Cognate compounds were poorly docked, and yet enrichment rates were high, e.g., factor Xa/MVP with an enrichment rate of 5.8.

Of these four possibilities, outcomes 1 and 3 are consistent with our expectations; if we get the docking pose correct, we would expect to enrich our ability to select active compounds from the pool of decoys. Outcome 2 implies a failure on the part of the scoring function; the scoring function did not differentiate between active and inactive compounds even though the docked poses being evaluated were largely correct. Outcome 4 is particularly troubling because it suggests that the enrichment rates are a result of chance; we are getting the right result for the wrong reason.

B.8. The application of knowledge about a protein target improved enrichment and consistency, and the application of such knowledge did not aid in the rapid identification of diverse chemotypes. We have demonstrated that while algorithm performance was inconsistent across the protein targets evaluated by this study, docking algorithms could enrich and identify leads. But are there ways to improve performance and consistency? One approach for improvement in enrichment could be found in the PDF data.

The *E. coli* and *S. pneumococcus* PDF targets are bacterial metalloproteases. Thus, it is reasonable to assume that inhibitors of these targets will contain metal binding moieties. The docking program Fred has the ability to use a SMARTS string defined pharmacophore constraint during docking. As is shown in Figure

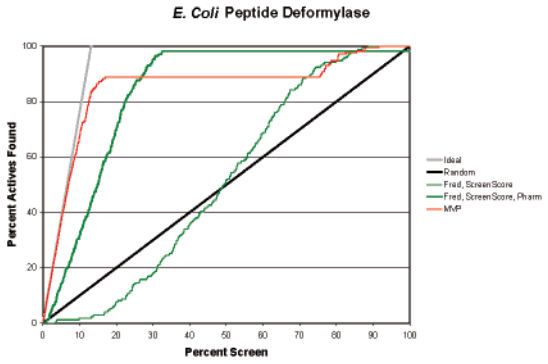

Figure 8. Boost plot demonstrating an improvement in enrichment by the docking program Fred when knowledge about *E. coli* PDF (heavy green line) is used versus no knowledge (light-green line). In this case, the only knowledge applied was a metal binding SMARTS pharmacophore constraint near the active site metal resulting in a 26-fold improvement in enrichment.

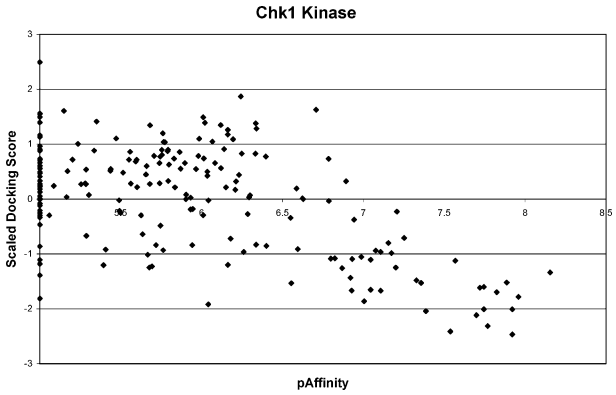

Figure 9. Plot of scaled score for Chk1 kinase vs pAffinity. The results from the scaled FlexX total score are depicted here. The correlation coefficient in this case is $r = -0.57$. The majority of the correlation comes not from the correct ordering of the compounds by affinity but from a low false negative rate.

8 for *E. coli* PDF, when Fred was run in a naive manner against either PDF target, its performance was at or worse than random. However if a metal binding pharmacophore constraint was placed near the metal, then enrichment for *E. coli* PDF and *S. pneumococcus* PDF increased by factors of 26 and 10, respectively.

One of the guidelines for this evaluation was complete freedom on the part of the program expert to optimize program performance. The docking algorithm MVP, as applied in this study, used atom-typed target points for superimposing ligand conformations within the protein binding site. These target points or pharmacophore points were determined manually by inspecting protein/ligand complexes from homologous protein structures. As a result, the target points contained more knowledge about binding for the protein target than similar nonatom typed target points did. For six of the eight targets evaluated, MVP met the enrichment criteria for success and for HCV polymerase had the highest enrichment rate of the programs evaluated for this target (Table 4). This result demonstrates that the application of knowledge about the target, gleaned either from the target itself or from homologous proteins, improved performance and consistency. However, the improvement in enrichment came with a cost. MVP was less successful than other algorithms at rapidly identifying all active chemotypes (Table 5). These examples demonstrate that the application of knowledge about a target or a compound class (informed docking) improved enrichment performance when compared with naive docking.

C. Scoring as an Affinity Prediction Tool. In the first results section, Prediction of Protein Bound Conformations, we have shown that docking algorithms could generate the experimentally observed small-molecule conformation and binding mode for a protein target. In other words, docking algorithms could essentially do virtual crystallography, although scoring functions could not reliably identify the best-docked pose. In the second section, Docking as a Virtual Screening Tool, we showed that, at a coarse level, scoring functions could distinguish active chemotypes from inactive chemotypes. While algorithm performance was inconsistent, with skilled use and application of knowledge about the protein target, docking algorithms and their associated scoring functions could perform virtual screening and identify leads. In this last section, we probe more finely the performance of scoring functions associated with docking algorithms and ask if these functions can distinguish between active and inactive molecules within a congeneric series or across several active series. In other words, can docking algorithms be used to predict potency or rank compounds by potency for lead optimization?

We point out that as part of our effort to evaluate docking algorithms under conditions similar to those used for lead optimization, the data sets used in this part of the evaluation are of moderate size and contained from two to five congeneric series spanning an activity range greater than 3.5 log units. This allows for an evaluation of algorithm performance for potency prediction within and across chemical series while reducing the likelihood of spurious correlations present in data sets of small size.

C.1. No strong correlation was observed for any scoring function protein target pair. Even a cursory examination of the data revealed that there is no statistically significant correlation between measured affinity and any of the scoring functions evaluated across all eight protein targets examined (Table 7). An extremely modest positive correlation was observed for Chk1 kinase target, with the largest correlation coefficient (r) observed being -0.57 (Figure 9). However, all of the correlation between measured affinity and docking score resided within a single compound series ($r = -0.64$). No correlation ($r = 0.0$) was observed for the second compound series directed toward this kinase target (Figure 10).

Table 7. Best Correlation Coefficient r between the $-$log Affinity (pAffinity) and Docking Score for All Programs across All Targets

program	Chk1	FXa	gyrase B	HCVP	MRS	*E. coli* PDF	*Strep* PDF	PPARδ
Dock4	−0.33	−0.31	−0.39	0.00	−0.13	−0.38	−0.34	0.07
DockIt	−0.49	−0.19	−0.37	0.04	−0.28	−0.13	−0.30	−0.34
FlexX	−0.57	−0.31	−0.39	−0.12	−0.01	−0.42	−0.25	−0.36
Flo+	−0.44	−0.38	−0.36	−0.09	0.05	−0.27	−0.39	−0.42
Fred	−0.14	0.01	−0.13	−0.07	0.13	0.07	−0.24	0.06
Glide	−0.47	−0.08	−0.21	−0.04	0.08	−0.13	−0.12	−0.35
Gold	−0.42	−0.05	−0.14	−0.09	0.04	−0.12	−0.11	−0.43
LigandFit	−0.45	−0.13	−0.39	−0.06	−0.15	−0.21	−0.49	−0.10
MOEDock	−0.29	0.00	0.07	−0.01	−0.13	0.08	0.20	0.17
MVP	−0.26	0.10	−0.33	−0.01	−0.18	−0.17	−0.16	−0.18

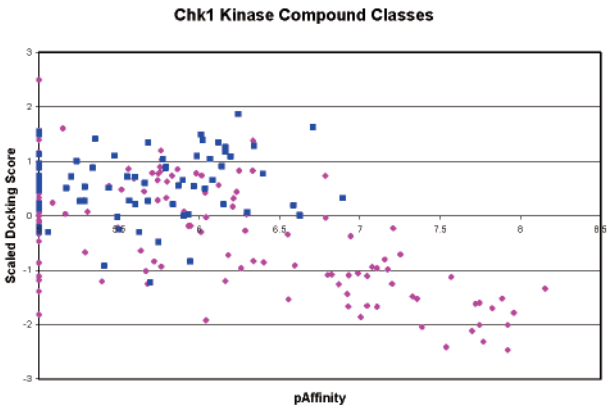

Figure 10. Plot of scaled score vs pAffinity where the two Chk1 kinase chemical classes are plotted in magenta (class **1**) and blue (class **2**). It is readily apparent that all of the correlation observed between the scaled docking score and affinity is found in the class **1** molecules and that no correlation exists between the docking score and class **2** compound affinities.

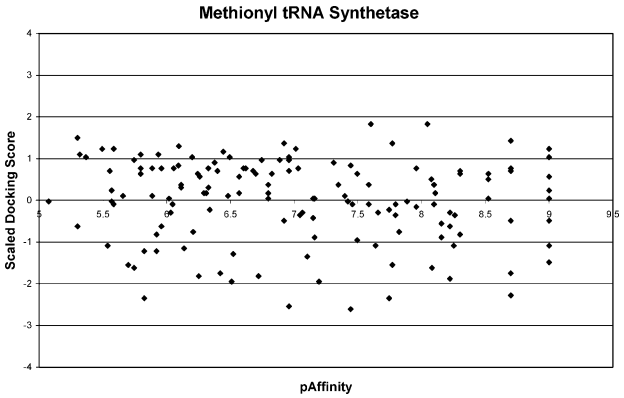

Figure 11. Plot of scaled score vs pAffinity for MRS and PPARδ. While the calculated correlation coefficient for the data shown for MRS is $r = -0.28$, this plot clearly demonstrates that these values are meaningless. No useful correlation exists between the docking score and compound affinity.

There were statistically insignificant correlations ($r = -0.5$ to -0.3) between affinity and docking scores observed for the gyrase B, factor Xa, PPARδ, PDF, and MRS targets (Table 7). We present a single illustrative example from the MRS data of the many pAffinity versus scaled score plots generated, but not shown, as part of our analysis of these data. Though the correlation coefficient calculated for the plotted MRS data is -0.3, even a superficial examination of Figure 11 reveals that no useful correlation existed between the measured affinity and the docking score. For HCV polymerase, no correlation ($r \leq -0.1$) between score and measured affinity was observed for any of the 37 scoring functions analyzed as part of this evaluation. The complete results are tabulated in Supporting Information.

The observed lack of a strong correlation between affinity and score for PDF, the metal-containing protease target in this study, was surprising because previously published data reported a strong correlation for peptidic inhibitors of human metalloproteases ($r^2 = 0.78$)[22] and for dicarboxylic acid inhibitors of metallo-β-lactamase ($r^2 = 0.87$).[25] It has been noted previously that success at potency prediction is more likely when the members of a congeneric series are of similar size and do not have large conformational differences between the protein bound and solution states.[35] The molecular weight range for each of the three PDF compound classes was greater than 180. One possible explanation for the contrast in correlation between affinity and docking score observed for this study versus previously published data could be the compound size variation present in this data set.

Table 8. Comparison of the Best Correlation Coefficient *r* between pAffinity and Docking Score versus the Correlation Coefficient between pAffinity and Score for Top-Ranked Poses with rmsd of ≤2 Å [a]

	MRS			PPARδ		
program	no. of well-docked ligands	all data	good pose	no. of well-docked ligands	all data	good pose
FlexX				17	−0.36	−0.56
Flo+				29	−0.42	−0.36
Glide	17	0.08	0.50	16	−0.35	−0.54
Gold	23	0.04	0.01	21	−0.43	−0.72
MVP				22	−0.18	−0.31

[a] The comparison is shown for selected docking programs on two targets, MRS and PPARδ.

A general observation with respect to scoring function performance on this data set is that no scoring function was able to rank-order within the congeneric series or to predict compound potency across series. Except for the case of *S. pneumococcus* PDF where the compound affinity was weighted toward nanomolar compounds, any correlation between docking score and affinity came from a reduction in the false negative rate (active compounds predicted to be inactive by the docking score) and not from a correct rank-order (data not shown).

C.2. In most cases, reproduction of the binding mode did not improve rank-order or potency prediction performance. For the targets included in this evaluation, no statistically significant correlation between docking score and affinity was observed. One possible explanation is that the docking algorithms did not reproduce the correct binding mode. According to this hypothesis, we would expect an improvement in correlation if the experimentally observed binding modes were evaluated by the scoring function. We remind the reader that for comparisons between pAffinity and scaled docking score, a correlation coefficient $r = -1$ would correspond to a perfect rank-ordering of compounds by affinity while $r = +1$ would mean that the scoring function was universally ranking poorly active compounds higher than more active compounds. Accordingly, we would hope that correlation coefficients would be more negative for well-docked compounds than for poorly docked compounds.

Two of the target data sets, PPARδ and MRS, contained a large enough number of cocrystal structures to allow us to assess whether affinity prediction improves for well-docked molecules. For each target, we computed a correlation coefficient for only those compounds for which the best-ranked pose was within 2 Å rmsd of the crystallographically determined pose. Table 8 lists the number of well-docked ligands for both of these targets along with correlation coefficients for the full data set and for the subset of well-docked ligands. Only programs that correctly docked at least 30% of the target-specific compounds are included in Table 8. The comparison between pAffinity and docking score for a single program is presented graphically in Figure 12. In this figure, all compounds in the data set are marked with diamonds while the well-docked compounds are emphasized by large squares.

Five programs were able to dock at least 30% of the cocrystallized PPARδ ligands within 2 Å of the crystallographically determined conformation (Table 8); the rest of the 54 cocrystallized ligands were poorly docked. For most of the compounds in the full PPARδ data set, we did not have

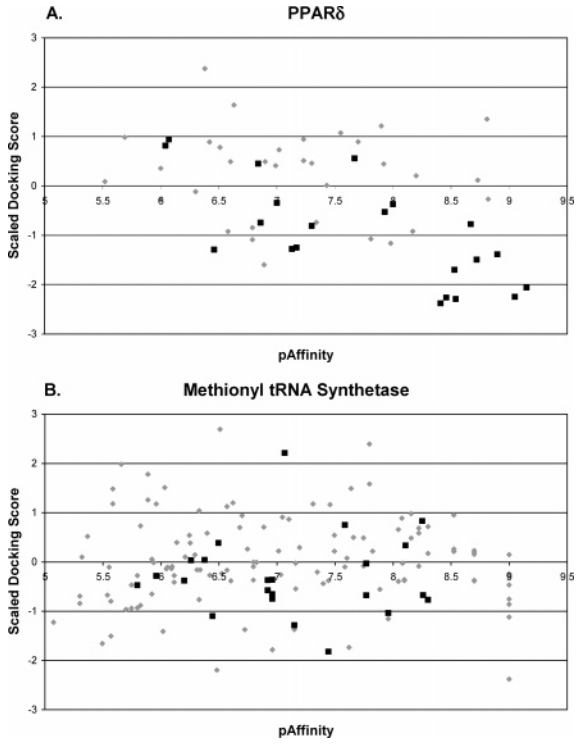

Figure 12. Plot of scaled score vs pAffinity for PPARδ (A) and MRS (B). Diamonds represent the score for the first pose returned by Gold, while the squares highlight scores where the rmsd is ≤ 2 Å from the crystallographically determined structure. For PPARδ we see a significant correlation between pAffinity and docking score for compounds known to be well docked while we see no correlation for well-docked MRS compounds.

crystallographic data to allow us to determine whether the compounds were well docked. For four of the five programs listed in Table 8, we saw an improvement in the correlation coefficient computed for only those compounds we know to be well docked. Indeed, for Gold the change was from a statistically insignificant $r = -0.43$ to a significant correlation $r = -0.72$. This result is depicted graphically in Figure 12A; the points marked by squares clearly show a trend in the right direction.

Conversely, only two programs were able to dock MRS compounds well, although in both cases >50% of the compounds were docked within 2 Å of the crystallographically determined conformation. In both of these cases, we saw no improvement in correlation between pAffinity and docking score for well-docked compounds. The absolute value of the correlation coefficient appears to have improved for Glide. However, $r = +0.5$ indicates that the scoring function had inverted predictions of affinity; less active compounds were being scored better than more active compounds.

In Figure 12 we compare the results generated by a single docking program, Gold, for both PPARδ and MRS. While we see a clear correlation between Gold docking score and pAffinity for PPARδ for well-docked compounds, we can see in Figure 12B that the points for well-docked compounds are distributed randomly throughout the graph. These differing results suggest that for PPARδ (but not for MRS) typical scoring functions applied to single well-docked poses captured key features contributing to binding affinity. This observation would be more compelling if the improvement in correlation were universally observed. However, even in the best case of PPARδ we saw only one example of statistically significant correlations between affinity and score. Indeed, for the program that docked more than 50% of the 54 cocrystallized PPARδ agonists within 2 Å of the crystal pose, we instead saw a slight decrease in correlation coefficient for well-docked compounds. More careful examination of systems such as PPARδ and MRS may therefore prove to be useful for guiding improvements to docking algorithms and scoring functions.

Conclusion

This evaluation has shown that docking programs are usually successful in generating multiple poses that include binding modes similar to the crystallographically determined bound structure. In the few cases where the reproduction of the observed binding mode was less precise, the difficulty was not in positioning the ligand within the binding site but in reproduction of the small-molecule conformation. In addition, we have shown that for the proteins used in this evaluation docking into a single protein structure by multiple compound classes did not prohibit reproduction of the observed binding mode even when the protein was held rigid. While docking programs were highly successful at reproducing binding modes, scoring functions were less successful at correctly identifying the binding mode. However, the decrease in performance was not extreme in that for five of the seven targets the success rate was greater than 40%. It is important to note that while some programs were consistently better than others at reproducing binding modes, no program was able to reproduce greater than 35% of the binding modes within 2 Å across all targets. This inconsistency highlights that while docking programs are being used successfully to predict binding modes, binding mode prediction is not a consistently solved problem and may require considerable intervention by a skilled computational chemist.

This evaluation of the performance of docking programs and scoring functions in the area of virtual screening has shown that active compounds could be found from a pool of biologically active decoy compounds. In most cases the detection rate by the top-performing algorithm was close to the theoretical limit through 5% of the score-ordered list. This result is even more compelling when one considers that for each of the targets at least 2% of the decoy compounds were inactive analogues of the active chemical series. Thus, we have shown that virtual screening can be successful when using data that mimic a typical pharmaceutical compound collection. While we have demonstrated that virtual screening is successful, we have also shown that in the absence of prior knowledge about the protein target program performance was inconsistent across the target types evaluated. This inconsistency means that when there is an absence of knowledge about the target, one cannot predict a priori whether a particular program will be successful against the given target. Because we have demonstrated that the application of knowledge about a target, whether broad or specific, could improve reliability in terms of enrichment, one obvious solution is to use all available knowledge when performing virtual screening. However, the application of knowledge comes with a cost, a loss in the diversity of the leads identified. The result of this behavior on the part of docking algorithms is that a burden is placed on the practitioner to make a determination early as to what is most important for virtual screening, fast and early lead detection or the identification of all diverse leads. One observation made by this evaluation was that high fidelity in the reproduction of observed binding modes did not automatically impart success in virtual screening. However, of particular concern was the observation that some scoring functions required no correct structural information for success in virtual screening. This result implies that under certain circumstances scoring functions are not ranking compounds based on structural information.

One goal of this evaluation was to quantify the relationship between docking scores and compound affinity. We have demonstrated that for the eight proteins of seven evolutionarily diverse target types studied in this evaluation, no statistically significant relationship existed between docking scores and ligand affinity. While this result was not unexpected given the large number of approximations used by docking scores to improve computation efficiency, it is the first time, to our knowledge, that an extensive evaluation of this area of docking and scoring has been published. We have shown in the binding mode prediction section that docking programs could reproduce the experimentally observed ligand binding modes. We have also shown that there was no consistent improvement in the correlation between docking scores and measured affinity if one considered only those docked poses similar to the crystallographically determined binding mode. Thus, good performance in reproduction of experimentally determined binding modes did not impart success in predicting affinity or in rank-ordering compounds by affinity within or across congeneric series. From the data generated by this evaluation, it is not clear what faults or failures cause poor ligand affinity predictions by docking scores because the performance was poor across all target types for all scoring methods tested. The fault was not in the reproduction of the experimentally observed binding mode by the docking programs or in the ability to differentiate active ligands from decoys but in the inability of the current scoring functions to distinguish, differentiate, and quantitate the sometimes subtle differences that can change ligand affinity from highly potent to inactive.

It was the goal of this docking and scoring evaluation to examine as systematically and exhaustively as possible the current state of the art in docking and scoring to determine where strengths and weaknesses exist. Docking programs were able to reproduce experimentally observed binding modes and in many cases identify that binding mode as the correct one. Docking programs and scoring functions could identify active ligands from a pool of decoy molecules. While it is true that greater reliability and accuracy in these two areas would be beneficial, the current state of the art allows for the useful application of these tools by a skilled computational chemist. However, in the area of rank-ordering or affinity prediction, reliance on a scoring function alone will not provide broadly reliable or useful information that can be applied to lead optimization. This study demonstrates unequivocally that significant improvements are needed before compound scoring by docking algorithms will routinely have a consistent and major impact on lead optimization. Because it is not completely obvious by what means these improvements will arise, it is our hope that studies such as this will provoke healthy debate and encourage significant collaborative research in the field.

Methods

A. Protein Targets. Protein structures were selected by system experts for each protein target. All hydrogen atoms were added to protein structures, and Asn, Gln, and His orientations were set automatically using the program Reduce.[36] In all but one case, hydrogens were added with no ligand present. The orientation and protonation of a His residue in PPARδ are affected by the presence of the negatively charged ligands, so in this one case, hydrogens were added with a representative ligand present. Apolar hydrogens were subsequently removed, and positions of polar hydrogen atoms were optimized under the CHARMm22 force field.

Initially, residues defining protein binding sites were identified using an automated procedure that located convex cavities on an α shape of the protein surface. These preliminary binding site definitions were then assessed and amended by each system expert. To define a common reference frame for all docking programs, the principal moments of inertia were computed for atoms of binding site residues; the geometric center of the binding site was placed at the origin, and principal axes were aligned along the *X*, *Y*, and *Z* axes.

By default, no crystallographic waters were included in binding site definitions. All crystallographic waters were oriented to match the protein orientation described above. Systems experts then provided commentary concerning the importance of specific waters for compound binding; each program expert made his/her own decision about how to incorporate that information in the docking calculations. Coordinates for protein and water atoms were provided to all program experts in PDB format along with a FASTA format sequence file identifying binding site residues.

All cocrystal structures for a given target were placed in the same orientation frame by a least-squares fit of backbone atoms to the protein structure used in docking calculations, and coordinates for ligand atoms were extracted. The rmsd calculations for predictions of bound conformation were carried out using SVL code provided by support scientists at the Chemical Computing Group. This code takes as input a database of ligands extracted from cocrystal structures and a database of docked poses, matches docked poses to the corresponding cocrystal structure, and computes the symmetry-corrected rmsd between the two. Volume integrals for computation of Tanimoto volume overlap T_{vol} were computed using the Shape toolkit from OpenEyes Scientific Software.

B. Ligand Set. Small-molecule ligands for each protein target were supplied as SMILES strings, and the full set of 1303 ligands merged into a single SMILES file. Random codes were assigned to all molecules in the merged set in order to ensure that compounds belonging to a given protein target and compound class were not clustered together in the input to the docking calculations. Chiral centers not explicitly denoted were identified and expanded using the Daylight tool Chiralify. For those compound classes where structural information on the absolute stereochemistry existed, the stereochemistry of congeneric compounds was restricted to the observed stereochemistry. All possible stereochemistries of chiral compounds for which stereochemical information was unavailable were generated and retained.

Initial three-dimensional conformations from the resulting SMILES file were generated using Corina version 1.8.1. These conformations were imported into MOE version 2002.03beta, and the compounds were ionized using the WashMDB function. Small-molecule conformations were minimized twice using the MMFF94 force field. During the first minimization hydrogen atoms were added, the initial conformations were rebuilt, distance-dependent electrostatics and nonbonded cutoffs were turned off, chirality was constrained to the initial chirality, the GBSA solvation model was turned off, and the structures were minimized to a gradient of 0.1. During the second minimization, structures were further minimized from the previous coordinate positions to a gradient of 0.01 with distance-dependent electrostatics turned on (dielectric 1, solvent dielectric 80, dielectric offset -0.09, $1-4$ scale 0.75, buffer 0.05), GBSA solvation model turned on, nonbonded cutoffs turned off, and chirality constrained to the initial chirality. Small-molecule conformations were visually inspected to ensure correct atom typing and hybridization states. The resulting coordinates were exported to SD files containing all hydrogen or only polar hydrogen.

C. Docking Programs. C.1. Dock4.[37−41] The initial ligand files in SD format were converted to Tripos mol2 format. Nonpolar hydrogen atoms were added to protein structures, Kollman 1994 charges were assigned to all protein atoms, Ni atoms were assigned a charge of $+2$, and protein atom coordinates and partial charges were saved in Tripos mol2 format.

To define the binding site for Dock4 calculations, all polar and nonpolar hydrogens were stripped from the protein, and the program dms as implemented in the Midas modeling package[42] was used to compute a molecular surface. For large systems, only residues within 8 Å (HCV polymerase) or 10 Å (factor Xa, MRS) of the predefined binding site were included in the molecular surface

calculation. Active site spheres were generated using SPHGEN[40] with default parameters. Sphere clusters were examined visually, and the cluster(s) that best filled the binding site as defined by the system expert were retained. The number of spheres selected ranged from 54 for *S. pneumococcus* PDF up to 148 for HCV polymerase.

Scoring grids were computed for a box extending 2−4 Å in all directions beyond the binding site sphere cluster(s). Grids were computed for the chemical, contact, and energy scoring methods. Parameters were set to default values with the following three exceptions: grid spacing set to 0.2 Å, use of an all-atom model, and use of a bump filter.

Three separate docking runs were carried out for each protein target, using chemical, contact, and energy scoring grids to drive the docking calculations. Flexible ligand docking was carried out for all molecules with 12 or fewer rotatable bonds. An anchor search using a simultaneous search method was carried out, and all docked ligands were minimized for 100 iterations. Ten docked poses were stored for each compound.

C.2. DockIt.[43] All docking calculations were carried out using DockIt, version 1.0. Protein coordinates were converted to CEX format using DockIt tools. Binding site spheres were generated by manually placing a dummy molecule in the binding site region and selecting the cluster overlapping with that dummy molecule. Where necessary, sphere parameters were adjusted to adequately represent the binding site as defined by the system expert. Ligand geometries were input, converted to tdt files, and converted to CEX format using DockIt tools. One-hundred docked poses were generated per ligand, all of which were output and rescored using the PLP[44] and PMF[45] scoring functions. In addition, two additional scores were calculated: DOCKPLP (the sum of DOCK and PLP scores) and DOCKPMF (the sum of the DOCK and PMF scores). The top-scoring 32 poses based on DockIt score were stored for each compound.

C.3. Flexx. All docking calculations were performed with FlexX, version 1.10.1,[46,47] as implemented in the version 6.8 release of the SYBYL modeling package.[48] Protonation states of binding site residues and torsion angles at the hydroxyl groups of serine, threonine, and tyrosine amino acids were set in the receptor description file by visual inspection of the PDB file with polar hydrogens.

Docking runs were carried out using the standard parameters of the program for iterative growing and subsequent scoring of FlexX poses. Two scoring functions as implemented within FlexX were used to score the poses. The default FlexX scoring function (a modified version of the empirical scoring function developed by Boehm[49]) and DrugScore (a knowledge-based scoring function[50]) were utilized for all docking calculations. Formal charges were used throughout all the simulations. Multiple conformations for rings were computed with the use of Confort.[51] The top 30 solutions were retained and stored in a single mol2 file. Finally, the saved poses were rescored by the following five scoring functions: Dock,[39] Gold,[52] PMF,[45] ChemScore,[53] and FlexX as implemented within CScore.[54]

C.4. Flo. All docking calculations were performed using Flo+, version 0802. The protein coordinates containing polar hydrogen were converted to Macromodel format using Flo+ tools. All residues within a 20 Å sphere centered around a residue identified visually as central in the binding site were selected, and the rest of the protein atoms were removed. The residues lining the binding site pocket (approximately 10 Å from same residue near the center of the active site) were selected to allow movement during minimization steps. The remaining residues were held rigid during all docking and minimization calculations. In the three cases (Chk1 kinase, factor Xa, and HCVP) where crystallographic water was present and included, the oxygen atom of the water was constrained, using a square-well potential, within a 0.2 Å sphere from its crystallographic position and attached hydrogen atoms were allowed to move freely during minimization.

Five docking algorithms present in Flo+, version 0802, were evaluated: mcdock (old scoring function), mcdock+, sdock+, fulldock+, and zdock+. The two mcdock algorithms rely on a Monte Carlo perturbation/fast search/energy minimization algorithm but use different scoring functions. For these algorithms 2000 steps of perturbation were performed and the 25 top-ranked poses retained. The remaining three methods are systematic docking algorithms with fulldock+ including 500 steps of a local Monte Carlo search followed by minimization to the results of sdock+. For sdock+ and fulldock+, the 25 top-ranked poses were retained. For zdock+, a single pose was retained.

For the virtual screening evaluation, three docking algorithms present in Flo+, version 0802, were evaluated: mcdock, mcdock+, and zdock+. The mcdock algorithms were run using 300 steps of perturbation, and the five top-ranked poses were retained. For the remaining method zdock+, a single pose was retained.

C.5. Fred. Ligand conformations were precomputed from the initial SD file using Omega, version 1.0. A maximum of 200 conformations were generated for each ligand, with an rms cutoff of 0.8 Å and an energy window of 8 kcal/mol. The maximum number of rotors was set to 30 to ensure that even the most flexible molecules in the set were included in the docking calculations.

Fred, version 1.2.1, docking calculations were carried out using protein structures with polar hydrogen atoms only and with the binding site definitions provided by system experts. Default Fred parameters were used with the following exceptions: the maximum number of poses passing through the shape-fitting filter was increased to 5000; κ was set to 1.75; γ was set to 0.0; the excluded volume was set to 5000 Å^3 for MRS and to 2500 Å^3 for all other protein targets. In separate docking runs, poses that passed the shape-fitting filter were scored using either ChemScore or ScreenScore. For PDF, additional docking runs were carried out using a pharmacophore filter to bias toward placing metal binding functionality near the Ni atom. For gyrase B, additional docking runs were carried out using a protein structure in which coordinates for a missing stretch of protein had been rebuilt using the homology modeling module in MOE. Up to 10 poses were saved for each docked ligand.

C.6. Glide.[55,56] All protein PDB files were minimized with Batchmin, version 8.0, in Maestro, version 2.0, using the MMFF94S force field, which promotes planarity of delocalized trigonal nitrogens, and using the water solvation model with extended cutoffs. All heavy atoms were constrained to their original PDB coordinates with a parabolic potential of 100 kJ/Å; 100 iterations of PRCG minimization were used in each case, which was sufficient to relax the hydrogen coordinates. Individual water molecules and metal ions were included as provided.

Binding sites were defined from the provided list of residue numbers using the ASL command language in Maestro. This was done to avoid biasing the site as a function of ligand scaffold. Glide grids were computed using these definitions for the inner grid box, which defines the range of motion for the center of each ligand. Outer (or enclosing) grid boxes were generally 15−20 Å longer than the inner grid box on each side, depending on the ligand length. The van der Waals (vdW) radii for nonpolar receptor atoms was scaled by 0.9.

Ligands were converted from .sdf format to .mae format using the Schrodinger utility sdconvert. Each ligand was then minimized with the MMFF94S force field (same as protein preparation) using the Schrodinger utility premin, which uses Batchmin version 8.0 and truncated Newton minimization and no solvation. Ligands were provided in ionized form where possible.

Docking calculations were performed with Glide (Impact, version 2.0) in standard sampling mode with maxkeep = 5000 and maxref = 400 and using the previously computed grids. The vdW radii for nonpolar ligand atoms was scaled by 0.8. Each docking job was run on a SGI server in parallel using the Schrodinger para_glide utility.

C.7. Gold. Gold, version 1.2, was used for all docking calculations. The ligand file bearing all hydrogens but with appropriately ionized polar groups was used for all Gold docking runs. GOLD atom-type checking was turned on for both ligand and protein atoms. For docking into HCV polymerase, two docking runs were carried out in which two key waters were either included explicitly

or omitted. For HCV polymerase, only the immediate active site was used to avoid known docking problems associated with the whole protein. For each protein, nonpolar hydrogens were added using Sybyl. In the case of PDF, which contains a bound Ni atom, Zn was substituted as the best surrogate for Ni.

Docking calculations were parallelized across the nodes of a Linux cluster, using Perl scripts to launch jobs and collate final results. The output was such that the directory hierarchy mimicked that which would be produced by a single GOLD run, facilitating the use of pre-existing Sybyl SPL scripts used for extracting the data into a Sybyl molecular spreadsheet. From this point, both the text file of GOLD energy data and an SD file containing docked ligand coordinates and the associated GOLD energy data were saved. The extraction SPL script clusters the results according to rmsd, and thus, only results that are ≥1.5 Å rmsd different from one another are represented. In practice, up to 50 different poses were saved for each docked ligand.

C.8. LigandFit. Parallel LigandFit was used as implemented in Cerius2, version ccN. Two independent docking runs were conducted for each protein target, one with the CFF, version 1.01, force field and the second with the Dreiding force field. Protein atoms were typed using the CFF or Dreiding force fields. For the PDF target, the Ni in the binding site was unbonded from amino acid side chains and typed as a Zn^{2+}. An SD file containing ionized ligands was used for docking. Ligands were autotyped and autocharged using the CFF and Dreiding force fields. The docking site was defined by using the Cerius2 site finder with a site opening size of 7.0 Å. The site was manually edited to include all of the binding site residues defined by the system expert. The energy grid was calculated with a distance-dependent dielectric constant of 1.0 and a nonbonded cutoff distance of 10.0 Å, and the grid was extended 5.0 Å from the site. Docking was performed with a flexible ligand. A variable number of Monte Carlo steps were used with the number of steps equal to 1000 times the number of torsion angles in the ligand. A Monte Carlo search step for torsions containing polar hydrogens was set to 30.0°. Site partitioning was used with three partitions. Rigid body minimization was performed on the four orientations of the docked ligand. Clustering of the docked ligands was performed with a maximum of 10 clusters per ligand and a rmsd threshold of 1.5 Å for cluster formation. Only diverse conformers were saved with a maximum of five conformers saved for each ligand. The docked ligands were scored with Ligscore2, PMF, and PLP1 scoring functions. For Ligscore2, the grid was extended from the site by 5.0 Å.

C.9. MOE. The standard docking routine as implemented in MOE, version 2002.03, was customized to enable high-throughput docking. Modified docking code was provided by the support scientists at CCG to run the algorithm in batch mode on a database of ligands. A routine was added to sort the database of docking results for each individual ligand by their total energy score. The best scoring orientation was then written out to a database that stored the optimal pose for each ligand in the test set. Docking of each ligand under standard stochastic search conditions was extremely time-consuming and not suitable for high-throughput mode. Consequently, the number of runs per ligand, the number of moves per run, and the length of the tabu list were reduced to a minimum to speed up the calculations. Additionally, the code was modified to reduce the number of random starting conformations of each ligand employed in the generation of the energy scoring grid. Finally, the predefined failure energy cutoff of the starting conformation was raised to 10^{12} kcal/mol to prevent premature termination of the docking run. The Engh−Huber united-atom force field implemented in MOE was employed for the docking calculations. MMFF94s parameters were used for the ligand, which was fully protonated. Partial charges were computed using the PEOE formalism as implemented in MOE.

The total interaction energy score returned by MOE includes an internal energy term without any reference to a low-energy conformation. To enable direct comparison of different ligands, the energy of each pose was recalculated using the MOE implementation of MMFF94s. A reference database of ligands was subjected to a limited stochastic conformational search, using MMFF94s, to generate a representative low-energy conformation of each molecule. The energy of this reference conformation was subtracted from the energy of the docked conformation and added to the vdW and electrostatic interaction energies between the ligand and protein to give a corrected docking score.

C.10. MVP. The MVP program[57] implements several different docking algorithms, including a "grow" procedure that grows ligands within the binding site and a "superdock" procedure that fits fully grown compounds into the binding site by superimposition onto target points. The growth procedure starts the growth process from an "anchor group" within each compound and works best when the binding orientation of the anchor group is known. The superdock procedure was used for this study to avoid any requirement for this prior knowledge. The superdock approach is broadly similar to that of the original DOCK program,[39] although the MVP implementation uses multiple atom types, a model for solvation, and more complete energy minimization. In addition, MVP accounts for desolvation and some aspects of configurational entropy by running two separate calculations for each compound, one in the binding site and one free in solution, calculating the binding energy using Boltzmann summations over the respective minima.[57] The superdock approach starts by using the grow procedure to run a conformational search calculation free in solution, retaining up to 1200 distinct low-energy conformations. Each of these conformations is then fitted into the binding site by superimposing four atoms or pseudoatoms from the ligand onto four target points of corresponding atom type within the binding site. The calculations used four main atom types: hydrogen bond donor, hydrogen bond acceptor, donor/acceptor (e.g., hydroxyl), and lipophilic. Target points were determined manually by inspection of available crystal structures, including protein/ligand complexes involving homologous protein structures. As with the DOCK program,[39] many different orientations are generated by using different matchings of the ligand atoms to the target points. These candidate orientations are initially refined with three steps of internal coordinate energy minimization with a short nonbonded interaction cutoff. Candidate orientations with sufficiently low energy were selected for six additional steps of internal coordinate energy minimization using a somewhat longer nonbonded cutoff. Orientations surviving this second filter were selected for 30 steps of Cartesian coordinate energy minimization with a longer nonbonded cutoff. Cluster analysis was used to identify redundant conformations in each cycle, effectively funneling the candidate binding modes down to a set of 50 low-energy, nonredudant binding modes. Throughout the calculation, energies are calculated with a simple solvation model based on solvent-exposed surface areas.[57] The initial conformational search used an additional term equal to −0.1 kcal per square angstrom of solvent-exposed area to penalize folded conformations and favor extended conformations. The binding energy is estimated as $E_{cpx} - E_{free}$, where the energies in the complex and free in solution are calculated from Boltzmann summations up to 50 conformations in the complex and up to 1200 conformations in solution, respectively, omitting the penalty term for folded conformations. This formulation captures desolvation effects and some portion of the configurational entropy of binding.

Acknowledgment. The authors acknowledge Drake S. Eggleston, Mike J. Corey, Colin M. Edge, Aldo G. Feriani, and Michael M. Hann for their support during this months-long multisite endeavor. We thank and acknowledge Ajita Bhat, Michael J. Bower, and Christine M. Richardson for their efforts in data collection. In addition, we thank Ajita Bhat and Michael J. Bower for their contribution, through discussion and comments, to the experimental design. We thank Dmitri Bondarev, Simona Cotesta, Sunny T. Hung, and Ryan M. Provencher for providing data analysis code. We especially acknowledge Hannah J. Shortley for her efforts in generating data from the program Gold. The authors thank the participants of an internal docking and scoring symposium, in particular Felix DeAnda,

Hideyuki Sato, Justin Caravella, and Marie-Hèléne Fouchet, for their contributions to discussions on the modification of the original experimental design and data analysis guideline development. Last, we thank and acknowledge a number of software vendors, in particular, Accelrys, Chemical Computing Group, OpenEye Scientific Software, Schrödinger, Thistlesoft, and Tripos, who provided free-of-charge demo versions of software, data analysis code, and/or technical support during this evaluation. Readers desiring more information about this evaluation and/or means by which access to the data may be granted are encouraged to contact the authors.

Supporting Information Available: Figures (pdf file) and spreadsheet files (Excel) showing the results of the docking programs for the targets. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) URL: http://www.rcsb.org/pdb/holdings.html.
(2) Knegtel, R. M. A.; Wagener, M. Efficacy and selectivity in flexible database docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 334−345.
(3) Ha, S.; Andreani, R.; Muegge, I. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 435−448.
(4) 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.
(5) Carlos, P.; Ortiz, A. R. Evaluation of docking functions for proteinligand docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.
(6) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(7) Doman, T. N.; McGovern, S. L.; Bryan, J.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.
(8) Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S. R.; Abagyan, R.; Samuels, H. H. Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7354−7359.
(9) Diller, D. J.; Li, R. Kinases, homology models, and high throughput docking. *J. Med. Chem.* **2003**, *46*, 4638−4647.
(10) Jenkins, J. L.; Kao, R. Y. T.; Shapiro, R. Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 81−93.
(11) 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.
(12) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(13) 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.
(14) 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.
(15) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(16) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J.-Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlén, M.; Stouten, P. F. W. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.
(17) 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.
(18) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.
(19) Makino, S.; Ewing Todd, J. A.; Kuntz, I. D. DREAM++: Flexible docking program for virtual combinatorial libraries, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(20) Esposito, E. X.; Baran, K.; Kelly, K.; Madura, J. D. Docking of sulfonamides to carbonic anhydrase II and IV. *J. Mol. Graphics Modell.* **2000**, *18*, 283−289.
(21) Tao, P.; Lai, L. Protein ligand docking based on empirical method for binding affinity estimation. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 429−446.
(22) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.
(23) 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.
(24) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(25) Olsen, L.; Pettersson, I.; Hemmingsen, L.; Adolph, H.-W.; Jørgensen, F. S. Docking and scoring of metallo-β-lactamases inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 287−302.
(26) 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.
(27) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.
(28) Zhou, B.-B. S.; Bartek, J. Targeting the checkpoint kinases: chemosensitization versus chemoprotection. *Nat. Rev. Cancer* **2004**, *4*, 216−225.
(29) Quan, M. L.; Smallheer, J. M. The race to an orally active factor Xa inhibitor: Recent advances. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 460−469.
(30) Gross, C. H.; Parsons, J. D.; Grossman, T. H.; Charifson, P. S.; Bellon, S.; Jernee, J.; Dwyer, M.; Chambers, S. P.; Markland, W.; Botfield, M.; Raybuck, S. A. Active-site residues of *Escherichia coli* DNA gyrase required in coupling ATP hydrolysis to DNA supercoiling and amino acid substitutions leading to novobiocin resistance. *Antibact. Agents Chemotherapy* **2003**, *47*, 1037−1046.
(31) Vaughan, M. D.; Sampson, P. B.; Honek, J. F. Methionine in and out of proteins: Targets for drug design. *Curr. Med. Chem.* **2002**, *9*, 385−409.
(32) Serre, L.; Verdon, G.; Choinowski, T.; Hervouet, N.; Risler, J.-L.; Zelwer, C. How methionyl-tRNA synthetase creates its amino acid recognition pocket upon L-methionine binding. *J. Mol. Biol.* **2001**, *306*, 863−876.
(33) Wang, M.; Ng, K. K.-S.; Cherney, M. M.; Chan, L.; Yannopoulos, C. G.; Bedard, J.; Morin, N.; Nguyen-Ba, N.; AlaouiIsmaili, M. H.; Bethell, R. C.; James, M. N. G. Non-nucleoside analogue inhibitors bind to an allosteric site on HCV NS5B polymerase. Crystal structures and mechanism of inhibition. *J. Biol. Chem.* **2003**, *278*, 9489−9495.
(34) Kliewer, S. A.; Xu, H. E.; Lambert, M. H.; Willson, T. M. Peroxisome proliferator-activated receptors: From genes to physiology. *Rec. Prog. Horm. Res.* **2001**, *56*, 239−263.
(35) Holloway, M. K.; et al. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995**, *38*, 305−317.
(36) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(37) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078−1082.
(38) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-based molecular design. *Acc. Chem. Res.* **1994**, *27*, 117−123.
(39) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, 1175−1189.
(40) 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*, 269−288.
(41) Ewing, T. J.; Makino, S.; Skillman, G. A.; Kuntz, I. D. Dock 4.0. Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, 411−428.
(42) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Robert, L. The MIDAS display system. *J. Mol. Graphics* **1988**, 13−27.
(43) Blaney, J. M.; Dixon, J. S. *DockIt*, version 1.0; Metaphorics, LLC: Mission Viejo, CA; www.metaphorics.com/products/dockit.html.
(44) Gehlhaar, D.; Verkhiver, G.; Reijto, P.; Sherman, C.; Fogel, D.; Fogel, L.; Freer, S. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.
(45) Muegge, I.; Martin, Y. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(46) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(47) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: Placing discrete water molecules during protein−ligand docking predictions. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 17−28.
(48) *Sybyl*, version 6.8; Tripos Inc., St. Louis, MO.
(49) Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(50) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(51) Balducci, R.; Pearlman, R. S. Confort: A Rational Conformation Analysis Tool. *Abstracts of Papers*, 217th National Meeting of the American Chemical Society; American Chemical Society: Washington, DC, 1999.
(52) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(53) 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.
(54) CScore as implemented in Sybyl version 6.8; Tripos Inc., St. Louis, MO.
(55) 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.
(56) 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**, 1750−1759.
(57) Lambert, M. H. Docking Conformationally Flexible Molecules into Protein Binding Sites. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Dekker: New York, 1997.

JM050362N