Lead Finder: An Approach To Improve Accuracy of Protein−Ligand Docking, Binding Energy Estimation, and Virtual Screening

Oleg V. Stroganov,[†] Fedor N. Novikov,[†] Viktor S. Stroylov,[†] Val Kulkov,[‡] and Ghermes G. Chilov*[,†]

MolTech Ltd., Leninskie gory, 1/75A, Moscow 119992, Russian Federation, and
BioMolTech Corp., 226 York Mills Road, Toronto, Ontario M2L 1L1, Canada

Received May 15, 2008

An innovative molecular docking algorithm and three specialized high accuracy scoring functions are introduced in the Lead Finder docking software. Lead Finder's algorithm for ligand docking combines the classical genetic algorithm with various local optimization procedures and resourceful exploitation of the knowledge generated during docking process. Lead Finder's scoring functions are based on a molecular mechanics functional which explicitly accounts for different types of energy contributions scaled with empiric coefficients to produce three scoring functions tailored for (a) accurate binding energy predictions; (b) correct energy-ranking of docked ligand poses; and (c) correct rank-ordering of active and inactive compounds in virtual screening experiments. The predicted values of the free energy of protein−ligand binding were benchmarked against a set of experimentally measured binding energies for 330 diverse protein−ligand complexes yielding rmsd of 1.50 kcal/mol. The accuracy of ligand docking was assessed on a set of 407 structures, which included almost all published test sets of the following programs: FlexX, Glide SP, Glide XP, Gold, LigandFit, MolDock, and Surflex. rmsd of 2 Å or less was observed for 80−96% of the structures in the test sets (80.0% on the Glide XP and FlexX test sets, 96.0% on the Surflex and MolDock test sets). The ability of Lead Finder to distinguish between active and inactive compounds during virtual screening experiments was benchmarked against 34 therapeutically relevant protein targets. Impressive enrichment factors were obtained for almost all of the targets with the average area under receiver operator curve being equal to 0.92.

INTRODUCTION

Molecular docking is a well established technique to model protein−ligand binding with numerous applications ranging from pure fundamental studies to industrial drug discovery research. While significant efforts have been invested into developing docking approaches over the past two decades, the existing molecular docking software still fails to satisfy the pragmatic expectations of researchers, especially of those from the industry. This is likely one of the main reasons for the sparse application of *in silico* ligand screening in the drug discovery process relative to experimental screening. Currently, the key directions to mastering molecular docking are 2-fold: a) search for a better representation of energy of intermolecular interactions (the so-called scoring functions) and b) develop docking algorithms that perform global optimization on multidimensional potential energy surfaces. As for the free energy calculations, though the main forces contributing to protein−ligand binding (i.e., van der Waals, electrostatic, hydrophobic, hydrogen bonds, and other specific interactions) are well-known, there are equally as many implementations of scoring functions as there are docking software programs, and the search for better representations persists.[1] The ideal scoring function must return the free energy of protein−ligand binding, and thus it must account for multiparticle changes associated with protein, ligand, and the surrounding aqueous medium upon binding. This task is not as straightforward as single-point energy calculations, which form the basis of molecular mechanics (and quantum chemistry as well); consequently, improving the accuracy of a scoring function is a serious scientific challenge. Another aspect of molecular docking performance, the docking algorithm, is related to the mathematical problem of global optimization. A number of modern mathematical approaches have been implemented in molecular docking software, such as Monte Carlo algorithms (ICM,[2] GlamDock[3]), genetic algorithms (AutoDock,[4,5] Gold,[6] MolDock[7]), incremental construction of an optimal ligand pose (DOCK,[8] FlexX,[9,10] Surflex[11]), systematic analysis of possible minima using graph searches (eHits[12,13]), and algorithms using hierarchical scoring functions for crude shape fitting and finer optimization of ligand pose (QXP,[14] LigandFit,[15] Glide[16,17]).

From the practical point of view, the quality of molecular docking methods is often characterized by a) docking success rate, which specifies the percentage of correctly predicted ligand positions in a given set of known protein−ligand complexes, and b) enrichment in virtual screening studies, in which ligands known to be active for a particular target protein are mixed with inactive ones (decoys) and rank-ordered by a docking program according to their predicted scores (or the binding energy). There are ongoing methodological disputes on making benchmarking studies more standard and meaningful. These include the following: the

* Corresponding author phone: +79163074842; fax: +74959394653; e-mail: Ghermes@moltech.ru.
[†] MolTech Ltd.
[‡] BioMolTech Corp.

10.1021/ci800166p CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/13/2008

proper definition of a docking success rate,[18] the construction of meaningful test sets of protein−ligand complexes,[19,20] data sharing,[21] the preparation of protein and ligand structures for modeling,[21] benchmarking virtual screening performance,[22−24] the construction of meaningful sets of active and decoy ligands for virtual screening experiments,[21,25,26] and reporting of the results of benchmarking studies.[21] However, the developers have not yet started applying a common set of standards in benchmarking studies. In particular, the attention paid to the accuracy of binding energy estimations provided by docking software has been less than adequate. Although some programs like AutoDock and Glide XP estimate the free energy of protein−ligand binding, it is commonly believed that binding energy prediction is a postdocking procedure and therefore more computationally demanding methods like free energy perturbation, linear interaction energy approximations, and others must be used for that purpose. Clearly, faster and more reliable estimations of binding energies are needed from docking programs.

In this article we introduce Lead Finder, a novel molecular docking software that provides scientists with fast and accurate ligand docking and binding energy predictions. Lead Finder combines the classical genetic algorithm with multilevel local optimizations. Energy calculations are performed using the original semiempiric molecular mechanics functional and are implemented in several discrete forms with varying speed/accuracy ratios. Lead Finder introduces three specialized scoring functions designed to rank predicted ligand poses, estimate the binding energy of docked ligand poses, and rank individual compounds in virtual screening experiments. The performance of Lead Finder has been benchmarked on the following: a) the binding energy predictions of over 330 protein−ligand complexes with experimentally measured binding energies; b) the docking success rate of 407 protein−ligand complexes borrowed from well-known test sets of such programs as FlexX, Gold, Glide, LigandFit, MolDock, and Surflex; and c) the virtual screening performance of 34 protein targets. The Experimental Section describes the following: a) the docking algorithm; b) the scoring functions; c) and the data sets for benchmarking free energy prediction, docking success rate, and the virtual screening performance. The Results and Discussion section contains a detailed presentation of our benchmarking study and a critical analysis of observations. The additional data that are crucial for reproduction and unbiased comparisons of the obtained results can be found in the Supporting Information.

EXPERIMENTAL SECTION

1. Overview of the Scoring Function. The Lead Finder approach for scoring function construction is based on a semiempiric molecular mechanical functional, which explicitly accounts for several different types of interactions described below. Individual energy contributions are scaled with empiric coefficients to produce three scoring functions fine-tuned for the following: (1) accurate binding energy prediction, (2) correct ranking of docked ligand poses, and (3) correct rank-ordering of active and inactive compounds in virtual screening experiments. These three specialized scoring functions are based on the same set of energy contributions with different energy-scaling coefficients.

The purpose of the first function, which we called the dG-scoring function, is to accurately estimate the free energy of ligand binding for a particular structure of a protein−ligand complex, The scaling coefficients for this type of scoring function have been derived by fitting calculated binding energies to the experimental values for a set of protein−ligand complexes with known 3D-structure and experimentally measured binding constants (as described below).

The purpose of the second scoring function, which we called the ranking scoring function, is to give the highest score to the correct (experimentally observed) ligand pose. Thus, the scaling coefficients of this scoring function have been fine-tuned to achieve the maximum docking success rate (maximum number of top-scoring poses with correct geometry) on a fixed set of protein−ligand complexes with known 3D-structures.

Finally, another set of scaling coefficients has been designed to yield maximum efficiency in virtual screening experiments. This third scoring function, called the virtual screening or VS-scoring function, assigns higher scores to active ligands (true binders) than to inactive ones.

The force field parameters used to construct the scoring functions were partially borrowed from the united-atom CHARMM19 field[27] and partially de novo optimized in order to increase the accuracy of docking and energy prediction. The united-atom ligand representation was chosen for two reasons: to reduce the number of parameters (compared to the full-atom force fields) and to speed up energy calculations. Bond stretching and valence angle bending were not considered because ligand bond lengths and valence angles were kept fixed during docking. The torsion energy was calculated in a different way as described below, so the torsion parameters were also dropped off. Grids for all types of interactions were calculated at a certain (user-adjustable) space stepping. The energy calculations during docking used a trilinear interpolation scheme to approximate grid value at a certain point in space by values in eight precalculated neighboring points.

The following energy terms corresponding to the different types of contributions to protein−ligand binding were used in the Lead Finder scoring functions.

van der Waals interactions are calculated with the 6−12 Lennard-Jones potential

$$\Delta G_{VdW} = k_{vdW} \sum_{i \in ligand} \sum_{j \in protein} LJ_{ij}(r_{ij}) \tag{1}$$

where k_{vdW} is a corresponding scaling coefficient. The summation runs over protein and ligand atoms except hydrogen bond acceptors and hydrogen atoms−for these atoms, the bonding energy of hydrogen is calculated instead. $LJ_{ij}(r_{ij})$ is a smoothed Lennard-Jones potential that is dependent on the types of atoms *i* and *j*. Parameters for calculating $LJ_{ij}(r_{ij})$ for standard atom types (O, N, C, H, CA − aromatic carbon, NX − nitrogen atom that cannot accept hydrogen bonds, P, S) were taken from the CHARMM19 force field, with some modifications introduced for a better representation of united atoms. The parameters for halogens were taken from OPLSAA and reoptimized. The standard Lennard-Jones potential produced poor results when protein−ligand overlapping was taking place and therefore was modified to smooth energy inside the protein interior and to mimic local protein flexibility by broadening energy minima.

Interactions with metals are also calculated with 10−12 Lennard-Jones potential in the form

$$\Delta G_{Me} = k_{Me} \sum_{\substack{i=liganddonoratom \\ j=metal}} \alpha_{i,j} LJ_{ij}(r_{ij}) \quad (2)$$

where k_{Me} is a scaling coefficient, a_{ij} depends on the metal coordination state and the relative orientation of ligand and metal orbitals, and LJ_{ij} is a smoothed 10−12 Lennard-Jones potential which accounts for the radial component of interaction energy. To calculate a_{ij}, the coordination state of a metal ion is detected first from a protein's 3D-structure, and the directions along which the coordination with ligand (O, N, S) atoms are possible are built up. Then a_{ij} is calculated for each vacant coordination direction as a 6th power of cosine of the corresponding angle between ligand−metal bond and the metal coordination vector. The smoothing of the Lennard-Jones potential is applied when ligand−metal overlapping takes place. The 10−12 potential was used instead of the 6−12 to catch the specificity of ligand−metal coordination. Then, the constants for the Lennard-Jones potential for metal−ligand interactions were adjusted after all other parameters of a scoring function had been set up. For that purpose, a set of approximately 100 protein−ligand complexes containing ligands coordinated with metal ions was extracted from PDB, and the parameters of LJ potential were adjusted to fit experimentally observed geometries. This way, the parametrization for Fe^{2+}, Fe^{3+}, Zn^{2+}, Mg^{2+}, Ca^{2+}, Mn^{2+}, metal ions and O, N, S ligand atoms was achieved. The energy-scaling coefficient (k_{Me}) was adjusted using the training set of protein−ligand complexes as described below.

Electrostatic interactions account for (a) the protein−ligand interaction itself and (b) the polar component of ligand desolvation upon binding. The protein−ligand electrostatic interactions were calculated using the screened Coulomb potential (SCP) with distance- and microenvironment-dependent dielectric permittivity. Following the original works of Mehler et al.[28,29] the energy of electrostatic interactions is calculated in Lead Finder as

$$\Delta G_{elec} = \sum_{i \in ligand} \sum_{j \in protein} k_{elec,n}(h_i, b_i) E_{elec,n}(h_i, b_i, r_{ij}, q_i, q_j) \quad (3)$$

where h_i denotes hydrophilicity of a microenvironment of atom i, b_i is its buried fraction, q_i and q_j are partial atomic charges, and r_{ij} is the interatomic distance. Hydrophilicity (h_i) of a microenvironment is a relative (compared to water) value that is calculated in accordance with the SCP electrostatic model.[28,29] The calculation of the atom's buried fraction (b_i) is described in the Supporting Information. The ligand's partial atomic charges were calculated using the Gasteiger algorithm.[30] Depending on h_i and b_i one of the three scaling constants ($k_{elec,0}$ $k_{elec,1}$, *or* $k_{elec,2}$), and one of the three functions to calculate electrostatic interaction energy ($E_{elec,0}$, $E_{elec,1}$ *or* $E_{elec,2}$) is chosen correspondingly. The calculation of $E_{elec,n}$ is described in the Supporting Information.

Furthermore, the electrostatic (polar) contribution of ligand desolvation upon binding to protein was evaluated using an adapted version of the Born model using the following formula

$$\Delta G_{ligand-born} = k_{lig-born} \sum_{i \in ligand} \frac{1}{2}\left(\frac{1}{D_{ES}(R_{B,i})} - \frac{1}{D_W(R_{B,i})}\right)\frac{q_i^2}{R_{B,i}} \quad (4)$$

where $D_{ES}(R_{Bi})$ denotes the dielectric screening calculated at the distance R_{Bi} (Born radius of atom i) from the center of ligand atom i in the protein−ligand complex, and $D_W(R_{Bi})$ denotes the dielectric screening calculated at the distance R_{Bi} in water. Born radii for different types of atoms were taken from publication[31] without further optimization. However, our findings suggested that the term $D_{ES}(R_{Bi})$ was quite sensitive to particular microenvironment characteristics. For that reason, additional parametrization of terms entering the screening function was performed to achieve better docking and scoring quality.

The hydrogen bonding energy contribution is calculated as a sum of energies of individual hydrogen bonds (or, briefly, H-bonds) formed between protein and ligand (E_{HB}) and energetic penalties arising from H-bond donors and acceptors in protein and ligand, which did not form H-bonds in the complex:

$$\Delta G_{HB} = k_{HB} E_{HB} + k_{HB,lig-pen} \Delta E_{HB,lig-pen} + k_{HB,prot-pen} \Delta E_{HB,prot-pen} + k_{HB,corr} N_{HB,corr} \quad (5)$$

The energy of individual H-bonds is calculated using the following formula

$$E_{HB} = \sum_{\substack{i \in ligand \\ j \in protein}} k(h_i) E_{HB,ij} \quad (6)$$

where coefficient $k(h_i)$ depends on the hydrophilicity of a particular H-bond microenvironment: for $h_i < -5$ one coefficient (for hydrophilic bonds) is taken; other bonds are treated as hydrophobic with another coefficient. The energy of an individual H-bond ($E_{HB,ij}$) is decomposed into angular and radial contributions according to the formula

$$E_{HB,ij} = c_{AHD,ij} \cdot c_{LP,ij} \cdot LJ_{ij} \quad (7)$$

where $C_{AHD,ij}$ is a squared cosine of an angle between acceptor atom, hydrogen, and donor atom; $C_{LP,ij}$ is a squared cosine of an angle between acceptor-hydrogen vector and acceptor-lone electron pair vector; and LJ_{ij} is a smoothed 10−12 Lennard-Jones potential.

The energy *penalties* for missing potential H-bonds in the protein−ligand complex were calculated by Lead Finder's own algorithm, which accounts for the accessibility of each H-bond donor and acceptor for water molecules and the strength of lost H-bonds upon ligand transfer from water to protein environment according to the following formula

$$\Delta E_{HB,lig-penalty} = \sum_{\substack{i \in ligand \\ i \in D,A}} (hb_p - f \cdot hb_w) \quad (8)$$

where hb_p and hb_w denote average numbers of H-bonds that the ligand forms in its protein-bound state and in its corresponding aqueous solution, and f is the degree of atom exposure to solution. The standard criteria (hydrogen-acceptor distance <2.5 Å, donor-hydrogen-acceptor angle >120°) were applied to count hb_p, while hb_w values were heuristically adjusted for different types of ligand atoms to maximize prediction accuracy.

The loss of protein hydrogen bonds induced by ligand binding was calculated by the Lead Finder's original method

$$\Delta E_{HB,prot-penalty} = \sum_{\substack{i\in ligand \\ i\notin D,A}} \sum_{j\in water} e^{-\frac{r_{ij}^2}{1.5}} \qquad (9)$$

according to which penalties are summed over all nonpolar ligand atoms overlapping with probable positions of water molecules hydrogen-bonded to protein polar atoms. The most probable positions of water molecules solvating protein in the ligand-unbound state are assumed to be 2.75 Å away from a hydrogen donor (acceptor) atom along the hydrogen-bonding direction (the direction spanned over lone electron pair). This penalty term was found to be crucial for the implicit accounting of protein specific desolvation arising from tightly bound water molecules.

The formation of correlated hydrogen bonds between hydrogen bond donors and acceptors that are separated by three or four chemical bonds is encouraged by the additional energy increment ($k_{HB},{}_{corr}$).

The nonpolar solvation favored by hydrophobic contacts in the protein−ligand complex was accounted for in a classical volume-based fashion[32]

$$\Delta G_{sol,V} = k_{sol} \sum_{\substack{i\in ligand \\ j\in protein}} S_i V_j e^{-r_{ij}^2/3.6} \qquad (10)$$

where summation runs over all protein and non-hydrogen ligand atoms, S_i and V_i denote atomic solvation parameters (energy increment and volume, correspondingly), and r_{ij} is the interatomic distance. It should be mentioned that the volume-based solvation term accounts primarily for nonspecific solvation effects, while more specific contributions are accounted for with additional terms calculated in a surface-based fashion according to the formula

$$\Delta G_{sol,S} = k_{polar-P} S_{L,polar-P} + k_{polar-S} S_{L,polar-S} + k_{nonpolar-P} S_{L,nonpolar-P} + k_{nonpolar-S} S_{L,nonpolar-S} \qquad (11)$$

where S denotes the area of contact (in Å^2) of polar or nonpolar ligand (L) atoms with protein (P) and solvent (S), and k are the corresponding scaling constants. Inclusion of the surface-based energy term reduces artifacts arising from the long-range and cumulative volume-based terms that often overestimate contributions from loosely bound (i.e., not forming direct contacts with protein) ligand moieties. Calculation of this energy term is computationally expensive and was therefore only considered in the most precise implementations of the scoring function (see ‘Types of Energy Calculations’).

The internal energy losses of the ligand upon transition from solvent to protein-bound state were accounted for by comparing the ligand internal energies in conformations typical for solution and protein-bound states

$$\Delta G_{\mathrm{int}\ ernal} = k_{nb}(E_{nb,ES} - E_{nb,W}) + k_{1-4}(E_{1-4,ES} - E_{1-4,W}) \qquad (12)$$

where k_{nb} and k_{1-4} are scaling constants for nonbonded (van der Waals) and 1−4 interactions (special case of nonbonded interactions between atoms separated by three chemical bonds). The first term in the sum is the difference of nonbonded energies of the ligand in protein−ligand complex and water, and the second term is the same difference of 1−4 interaction energies. Nonbonded interactions were calculated using the standard 6−12 Lennard-Jones potential without smoothing. 1−4 interactions were also calculated using the 6−12 Lennard-Jones potential but with the atomic radii reduced by 0.2 A. The special term for 1−4 interactions compensates for the absence of a fully functional description of torsion potential in the current implementation of Lead Finder. The direct inclusion of torsional penalties based on standard molecular mechanical torsional potentials is currently included only for a set of particular chemical bonds, such as the conjugated double bonds, conjugated aromatic bonds, OH-group adjacent to an aromatic or double bond, and carbonyl group adjacent to a double bond.

The torsion (dihedral) energy is calculated according to the formula

$$\Delta G_{dihedral} = k_{dihedral} \sum_{i\in dihedral} (V_{oi} + 0.5[V_{1i}(1+\cos a_i) + V_{2i}(1-\cos 2a_i) + V_{3i}(1+\cos 3a_i)]) \qquad (13)$$

where $\alpha_i = \varphi_i - \varphi_{0i}$ is the difference between the current and the equilibrium value of the *i*-th dihedral angle, and corresponding constants V are taken from the CHARMM19 force field. This energy term is calculated only for conjugated double and aromatic bonds.

Entropic losses accounting for freezing ligand’s degrees of freedom upon binding to protein are calculated in a standard linear fashion

$$\Delta G_{entrop} = k_{tors} n_{tors} \qquad (14)$$

where n_{tors} denotes the number of freely rotatable bonds (FRBs) in the ligand, except terminal groups consisting of a single heavy atom and attached hydrogen atoms. The internal rotation of such groups is believed to be preserved upon ligand binding. k_{tors} is a corresponding scaling factor which was fitted using the training set.

2. Overview of the Docking Algorithm. *Ligand Representation.* Following the formalism of genetic algorithms the ligand is represented by a chromosome, each gene of which codes a certain degree of freedom sampled during a docking run (for more details see the Supporting Information). In addition to common (translational, rotational, and torsional) degrees of freedom the conformations of flexible (five- and six-member) ligand rings are also stored in the chromosome. Ring conformations are generated using the ring-flapping algorithm described elsewhere.[33]

Energy Calculations. As mentioned above, Lead Finder provides three specialized scoring functions (ranking and dG- and VS-scoring functions) that use the same set of energy contributions albeit with different weighting factors. In addition, Lead Finder uses several simplified implementations of each scoring function at different stages of the docking process. Expanding sampling space at the initial stage of docking process is more significant to the final outcome than increasing the precision of calculations. At the final stage of docking process precision becomes more important in the selection of the most promising solutions (ligand poses).

Pose Optimizations. The local pose optimization is viewed as a valuable component[4] of genetic algorithm, which facilitates faster evolution of individuals (ligand poses) in addition to usual genetic operations such as recombination and mutation. Similar to energy calculations, a number of pose optimization algorithms are applied during a docking run to achieve the optimum balance between speed and accuracy. The pseudo Solis-Wets (PSW) optimization used in Lead Finder is based on a random displacement in each degree of freedom and following the chosen direction when the energy of a new ligand pose is lower. If no improvement in energy is detected, after a series of trials the step size is reduced, and trials are repeated until the step size reaches a specified limit. This version of PSW optimization is referred to here as "complete". In the "fast" version of PSW optimization, the optimization cycle stops after a specified number of trials when no sizable improvement in energy is detected. The "complete" pose optimization is used more frequently during the initial pool generation stage and less frequently at the later stages of the docking process.

A special novel algorithm is applied to find the closest local energy minimum for a given pose. This algorithm marks degrees of freedom (DOF) displacements along which there are gains in energy. A further PSW-like optimization proceeds in a subspace spanned along these degrees of freedom. After a specified number of failed trials the subspace is reduced to those DOF that gave improvement in the previous productive step. Optimization and reduction of subspace are continued until only one DOF is left; scanning the last DOF returns the local energy minimum.

The "most complete" local optimization algorithm is designed to scan local minima surrounding a given pose and to select the optimal one. This is achieved by scanning subspaces spanned over the single and double terminal DOF (i.e., the rotatable bonds attached to terminal rigid fragments, and fragments attached to terminal ones). Smoothed energy profiles are built for each subspace, and then for each mapped minima a local optimization is done as described in the previous paragraph. Depending on the discreteness of the built energy profiles, more or less precise optimization is achieved. The "most precise" method is used in postdocking pose optimization.

Initial Pool Generation. The input pool of individuals for docking is not chosen arbitrarily. It is in fact preoptimized to speed up *in silico* evolution. Before the start of a docking process, the conformation of a ligand in solution is optimized. Its energy is then used as a reference in energy calculations, and its structure is used as a source for generating the initial pool of individuals. The pool of structures is generated by randomizing the translation and orientation coordinates of a molecule.

Subsequently, a pool of 10,000 to 100,000 random individuals is generated, each of which is subjected to a set of optimization procedures. First, all structures are minimized with the "complete" pose optimization algorithm. Then, structures demonstrating better binding of some fragments are optimized with the "most complete" algorithm. Ligand poses are then energy-sorted and clustered by geometry. The input pool for the genetic algorithm is filled first by the best individuals from found clusters and then by other individuals according to their energy values.

Genetic Algorithm. The overall implementation of genetic algorithm is done as follows:

1. an initial pool (population) of individuals is generated.

2. until the convergence of population is achieved:

a. common operators (crossover, mutations, optimizations) are applied.

b. specific operators (crossover with individuals from elite niches, etc.) are applied.

c. individuals are divided into niches.

d. individuals are selected for further rounds of evolution.

3. docked poses are optimized and ranked and their dG-score and VS-score are calculated.

The initial pool generation and optimization are described in the above paragraphs. The crossover is two-point (spans between two chosen genes). Mutations are generated with Cauchy distribution. The probability of mutation depends on the stage of docking process. At the beginning, it is relatively low at 0.05, and grows up to 0.5 as the population matures. The number of offspring for a given individual is exponential to its rank. Additionally, the three worst individuals are subjected to randomization followed by the "fast" PSW optimization. The best individual in a niche is subjected to the "complete" PSW optimization when it is changed.

The current implementation of the genetic algorithm uses the notion of a niche to cluster individuals with similar genotypes and to restrict their expansion. A niche is represented by individuals whose genetic distance as defined by their weighted difference in gene values from the best individual (by ranking score) is less than a specified value. Niche size (the number of individuals in a niche) is then restricted. Thus, when new individuals are generated, their selection is preceded by clustering the population into niches. Then, selection is performed by sorting offspring according to their score and filling up niches. If the best individual of the best niche remains unchanged for a certain number of selection rounds, it is transferred to the list of elite niches and other individuals from that niche are automatically erased. Elite individuals do not participate in docking directly; however, they can form descendants with individuals from the current population. All individuals within a specified genetic distance from any one of the elite individual are automatically removed from the population; however, when an individual has lower energy than the elite one, the latter is replaced by the former.

After convergence of the population ligand poses are optimized using the most precise form of the scoring function as described above and the "most complete" optimization algorithm. Obtained poses are ranked, and the dG of binding is calculated.

3. Test Set for Benchmarking Binding Energy Calculations. A set of 330 protein−ligand complexes with experimentally measured binding constants and available 3D structures (from PDB) was chosen from the following databases: AffinityDB,[34] PDBbind,[35,36] BindingDB.[37] Structures for the test set were manually selected by a number of qualitative criteria: a) binding constant is present in at least two different sources (databases or scientific publications) and b) compounds with varying physicochemical properties (number of freely rotatable bonds, hydrophobicity (cLogP), molecular weight, charge, binding constant) were selected over compounds with similar properties in order to increase diversity of the test set. Out of the 330 complexes selected, 100 were used as a training set to parametrize dG-scoring function of Lead Finder. The other 230 compounds were used as a test set for an independent benchmarking of accuracy of binding energy prediction. Prior to docking, protein−ligand complexes were prepared (preprocessed) automatically as described below.

4. Test Sets for Benchmarking Docking Success Rate. The following sets of publicly available protein−ligand complexes were used to benchmark the docking success rate of Lead Finder: the original test set of Gold,[6,38] containing

Table 1. Average Number of Hydrogen Bonds Formed by Different Types of Ligand Atoms in Solution

donor/acceptor type	number of H-bonds
H (polar)	0.6
N (sp^2 or sp^3 hybrid)	0.6
O (sp^3 or sp^3 hybrid)	0.8

134 structures; the so-called 'clean' CCDC/Astex test set containing 92 structures with resolution of 2 Å and better;[19,39] the test set of Glide SP containing 282 structures[16] and Glide XP containing 268 structures;[17] the test set of Surflex containing 81 structures;[11] the test set of FlexX containing 200 structures;[10] the test set of MolDock containing 77 structures;[7] the set of 88 protein−ligand complexes used for benchmarking active site cavity detection (75 structures) and docking (19 structures) by LigandFit;[15] and the novel test set of Astex containing 85 carefully chosen diverse protein−ligand complexes.[20]

5. Protein Targets for Benchmarking Virtual Screening. In total, 34 protein structures listed below in Table 8 were selected for the assessment of Lead Finder performance in virtual screening experiments. These structures were chosen on the basis of a) protein relevance for drug discovery research; b) availability of high-quality 3D structure of a protein; and c) availability of a sufficient number of well-characterized active ligands. The availability of results of virtual screening studies obtained through other docking programs was also desirable in choosing a protein target for the current study. Finally, for each target protein a set of active ligands was extracted from the following public sources: PDB database,[40] KiBank,[41] and active ligands exposed by Surflex developers.[42] The complete list of selected active ligands for each target protein can be found in the Supporting Information where ligand structures are also provided in InChi and Smiles formats. The total numbers of selected active compounds per protein target are given in Table 8. A set of 1904 decoy ligands used in all virtual screening studies in the current work was borrowed from the original Surflex publication[43] that is available for downloads from the Surflex developers site.[42] We chose this set of decoys because of its easy availability and the meaningful criteria (such as diversity and drug-likeness) applied during construction of the set by the authors.[43]

6. Automatic Preparation of Proteins and Ligands for Docking. The protein and ligand structures were prepared for docking experiments automatically from source PDB files using the "build_model" application that is part of Lead Finder package. During the structure preparation process, hydrogen atoms were added to protein residues and ligands according to their predicted ionization state. The positions of functional hydrogen atoms were energy-optimized. The coordinates of heavy protein atoms were left unchanged. The detailed description of the algorithm implemented by Lead Finder to optimize the ionization states of protein residues and hydrogen bonding network will be published soon.[44] An evaluation copy of the "build_model" application is available from the authors upon request. The prepared protein structures used in the current benchmarking study can be freely downloaded from the developer's site.[45] Further details about the protocol of protein structure preparation can be found in the Supporting Information.

7. Docking Settings. Lead Finder allows the user to choose between the two types of docking regimes. The first one is faster and is called the screening regime. The other one is slower but more precise and is called the docking regime. When an end user chooses one of the two regimes, Lead Finder automatically determines the most appropriate configuration settings for the docking algorithm in order to achieve the optimal balance between speed and accuracy of docking calculations. The screening regime was designed to run docking as fast as possible while not significantly compromising the docking success rate and accuracy of binding energy estimations. The screening regime is therefore more suitable in virtual screening studies where the speed of calculations and correct energy ranking are both crucially important. The docking regime was designed to achieve the best docking success rate within a reasonable time frame, which we assumed to be 1 min of CPU time per compound on average. The availability of two preconfigured docking regimes also facilitates more thorough and more reproducible comparisons of Lead Finder with other programs. To compare the capabilities of these two regimes, all docking success rate and binding energy benchmarking studies were carried out in both screening and docking regimes. Virtual screening studies were carried out in the screening regime only.

To enable a fair comparison with other programs tested elsewhere we tried to bring the parameters of our test environment in line with those of other programs as much as possible. The energy grid was defined as a box spanning 6 Å in each Cartesian direction from the reference ligand coordinates. The reference ligand was always taken from the original PDB structure. As for the ligands, all five and six member cycles were treated as flexible. Sulfamide bonds were considered rotatable. Bonds connecting aromatic rings and double bonds were considered flexible, while amide bonds − rigid. Covalently bound ligands were treated in the same way as all other ligands.

8. Docking Success Rate Measurements and Comparative Studies. Although a number of sophisticated techniques are available to benchmark docking success rate, one particular method has been the most widely used where one defines the successfully docked ligand as a top-scored pose with rmsd from the reference ligand coordinates of less than 2 Å. In this study we stick to this definition of docking success to make Lead Finder benchmarks comparable with the competitive software benchmarks obtained elsewhere. Additionally, for the sake of reproducibility and statistical significance of the obtained results all docking calculations were performed independently 20 times, and only when the probability of generating a top-ranked pose within the 2 Å rmsd accuracy is greater or equals 0.5 (in other words, successful docking was obtained in 10 or more tests out of 20), docking was recognized successful. The probability check has not been widely used in docking success benchmarking. We are aware of its application only in Surflex[11] and Gold[20] benchmarking studies. However, we hope this additional testing will become more widely accepted in the future.[21]

In this article we also compare docking success rates achieved by Lead Finder on particular test sets of protein−ligand complexes (described above) with the competitive software benchmarks obtained on the same test sets. Information on the competitive software performance from the

Table 2. Distribution of Physicochemical Properties of Ligands and the Error of Binding Energy Estimation for 100 Complexes Used To Train the dG-Scoring Function (I) and the Entire Set of 330 Complexes (II)

Mw			cLogP			dHB			dHB+aHB			charge			dGexp-dGcalc, kcal/mol		
range	I	II	range	I	II	range	I	II	range	I	II	range	I	II	range	I	II
0 − 100	2	4	<0	19	67	0	12	30	0 − 2	12	35	<−4	0	5	−3 − −2	6	41
100 − 200	23	70	0 − 1	4	24	1	13	48	3 − 4	17	46	−4 − −2	11	42	−2 − −1	13	46
200 − 300	16	66	1 − 2	10	34	2	13	60	5 − 6	24	66	−2 − 0	59	199	−1 − 0	31	80
300 − 400	22	74	2 − 3	5	21	3	21	60	7 − 8	17	55	0 − 2	30	83	0 − 1	31	86
400 − 500	17	59	3 − 4	8	25	4	15	49	9 − 10	10	39	>2	0	1	1 − 2	13	50
>500	20	57	4 − 5	10	26	5	16	42	>10	20	89				2 − 3	6	27
			>5	10	31	>5	10	41									

Table 3. Docking Success Rates (%) of Different Programs Obtained on Their Native Test Sets and the Corresponding Lead Finder Success Rates in Docking and Screening Regimes

	FlexX[10]	Glide SP[16]	Glide XP[17]	Gold[6]	Gold[20]	Gold[19]	LigandFit[15]	MolDock[7]	Surflex[11]	all test sets
original data	46.5	70.2	69.4	72.4	76.5	n/a	n/a	87.0	70.4	n/a
Lead Finder (docking regime)	85.0	82.3	81.3	87.3	90.6	92.4	87.3	96.1	96.3	85.0
Lead Finder (screening regime)	76.5	77.3	77.2	81.3	78.8	83.7	82.3	79.2	76.5	79.0
number of structures	200	282	268	134	85	92	88	77	81	407

original publications is provided in Table 3 below. See the Supporting Information for additional comments.

9. Virtual Screening Benchmarks. For each out of the 34 selected target proteins virtual screening benchmarking consisted of creating a test library of compounds obtained by mixing active target ligands with a set of decoy compounds, docking each compound from the library to the target, and rank-ordering compounds according to the calculated VS-scores (not the dG-scores!). For quantitative characterization of virtual screening efficiency, the following parameters were used: the area under the so-called receiver operating curve (ROC), the enrichment factor (EF), and the true positive (TP) rate. The area under ROC plot (or simply ROC value or ROC) is an integral parameter of the virtual screening performance that corresponds to the area under the curve built according to the following rule: for a given fraction of the screened library the Y-coordinate denotes the fraction of active compounds found (true positives), and the X-coordinate represents the fraction of inactive ligand (decoy) compounds found (false positives). An ideal curve reflects 100% of true actives found and 0% of decoys; this ideal curve returns ROC = 1. The enrichment factor, on the contrary, is not an integral parameter. It is calculated for a certain percentage of active compounds as the fraction of active compounds found divided by the fraction of the screened library. For example, EF70 denotes the enrichment factor at 70% of active ligands found. In the current paper we provide comparative data for several enrichment factors (EF20, EF40, EF70). The true positive rate is defined as the fraction of active compounds recovered at a certain percentage of decoys. For example, the TP value of 0.5 denotes the fraction of active compounds recovered at 0.5% of recovered decoys. ROC values obtained in the current study are provided with confidential intervals as suggested in refs 21 and 46. The detailed enrichment plots for each target can be found in the Supporting Information. A special mention should be made regarding the way we calculated fractions of the whole library of compounds and decoy compounds: since the Surflex decoy set contained 1904 structures, some of which were represented in multiple conformations (differing by the conformation of flexible rings), we docked all 1904 structures independently but used only the best positions of 1100 unique compounds for enrichment factor calculations.

RESULTS AND DISCUSSION

1. Pose Ranking and Binding Energy Prediction. Protein−ligand complexes for both the training and the test sets were chosen on the basis of maximum diversity of ligand's physicochemical properties, such as molecular weight (Mw), cLogP, number of hydrogen bond donors (dHB) and acceptors (aHB), net charge, and wide range of protein−ligand binding energies. The availability of at least two independent references for each experimental constant was considered a plus in selecting data for training and test sets. On the basis of those criteria, 330 protein−ligand complexes (100 for the training set and 230 for the test set with no overlap between the two sets) were manually selected for parametrization and assessment of the binding energy prediction. As Table 2 shows, the physicochemical properties of selected ligands provided a substantially diverse representation of a chemical subspace satisfying the Lipinsky's Rule of Five.[47] However, ligands spanning beyond this subspace (57 compounds with Mw > 500, 31 compounds with cLogP > 5, 41 compounds with dHB > 5, 89 compounds with (dHB+aHB)>10) were also generously represented, which is important in view of the recent findings that physicochemical properties of drugs are not necessarily restricted by the Lipinsky rule.[48] As such, Lead Finder has been trained and tested to estimate binding energy in a broadly defined chemical space.

Our training methodology was different from that of other docking programs as Lead Finder implements distinct scoring functions for pose ranking and binding energy estimation and uses different data sets for training its scoring functions. 100 protein−ligand complexes out of the set of 330 complexes described above were used to train Lead Finder's dG scoring function. 407 complexes were used to train the ranking scoring function using ligand coordinates from their respective PDB complexes. On the contrary, Glide XP,[17] AutoDock 3.05,[4] or AutoDock 4[5] use a single scoring function for both purposes. In particular, calibration of the Glide XP scoring function was performed for both pose ranking and dG simultaneously using a set of 268 protein−ligand complexes, where only 198 of them had experimentally measured binding constants. AutoDock 3.05 and AutoDock 4 also used ligand coordinates from PDB complexes

Table 4. Docking Failures (In the Docking Regime) Attributed to Different Scoring Errors (E − Evident, M − Minor, D − Dubious Errors)[a]

PDB code	NFRB	rmsd, Å	dG ref, kcal/mol	error type
1acl	11	4.81	−7.0	D
1baf	6	7.09	−6.3	M
1bkm	16	2.73	−12.9	E
1c8k	5	5.78	−7.8	E
1d3d	10	2.97	−11.8	M
1d3p	12	2.81	−10.9	D
1e5i	4	4.69	−6.9	M
1elb	13	5.21	−8.1	E
1elc	13	7.01	−7.0	E
1fjs	7	2.33	−11.8	D
1g45	4	2.22	−7.8	D
1g46	4	4.69	−7.8	D
1g48	4	2.84	−8.8	D
1g4o	3	4.20	−8.4	D
1g52	4	4.20	−9.3	D
1g53	4	4.82	−8.7	D
1g54	4	5.15	−10.0	D
1gpk	0	3.59	−7.5	E
1htf	13	8.43	−11.4	E
1icn	15	6.93	−7.6	E
1n2v	3	2.35	−7.3	E
1pso	24	13.14	−13.0	E
1tlp	15	7.66	−11.4	D
1xm6	5	2.53	−7.9	D
2er6	30	10.71	−13.6	E
2plv	18	8.46	−8.0	E
3cla	8	7.55	−6.4	E
1mfe	15	6.05	−7.8	M
1n2j	3	3.33	−4.7	M
1sq5	6	4.31	−5.5	M
5tim	2	3.00	−5.7	E

[a] The PDB codes that were docked correctly in general (e.g. the ligand's pharmacophore group was correctly placed) are highlighted in bold; the underlined PDB codes correspond to unspecific protein-ligand complexes. The binding energies (dG ref) for the optimized reference ligand pose are provided.

to compute energy-scaling coefficients for their scoring function; however, the same scoring function was also used for pose ranking.

In our experiments with the scaling coefficients we found that some manual tweaks are necessary not only to minimize the rmsd between predicted and experimentally measured binding energies but also to obtain more realistic ratios between the different types of energy contributions. The necessity of introducing additional (even quite heuristic) considerations becomes obvious when we note the fact that most PDB complexes represent binding ligands; however, in real life we have to deal with nonbinders. The resulting energy scaling coefficients tailored for the dG- and VS-scoring functions are provided in Table 7.

The rmsd between calculated binding energies and experimentally determined values was found to be 1.24 kcal/mol for the training set of 100 complexes, 1.60 kcal/mol for the test set of 230 complexes, and 1.50 kcal/mol for the entire set of 330 protein−ligand complexes. To illustrate the quality of the Lead Finder dG-scoring function, error values (defined as the difference between predicted and experimental values) in binding energy prediction were plotted against the experimentally measured values (Figure 1). As shown in Figure 1 and Table 2, the distribution of error values in binding energy prediction is fairly symmetric for the test set. For the training set, it is symmetric by design, since the optimization of energy-scaling coefficients eliminates the constant error term.

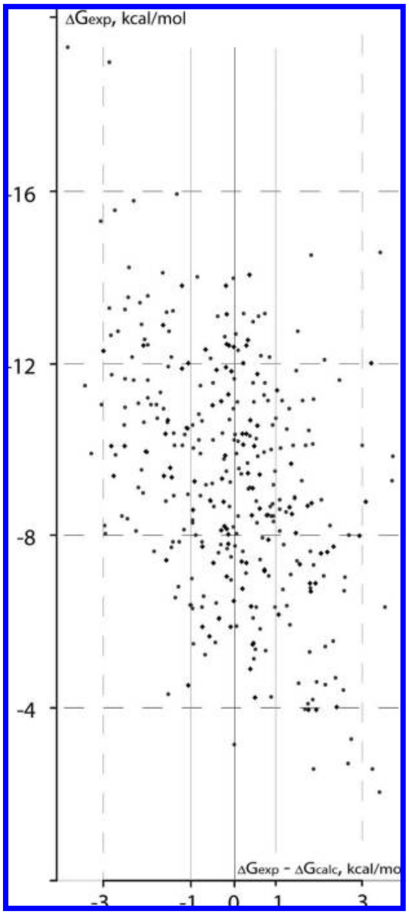

Figure 1. Binding energy prediction errors for the set of 330 protein−ligand complexes plotted against experimentally measured binding energy. Black dots correspond to the training set, gray − to the test set complexes.

About 50% of all protein−ligand complexes we tested fell into the ±1 kcal/mol bin and 79% fell into the ±2 kcal/mol bin of dG prediction accuracy. Unfortunately, the majority of available docking programs lack the functionality of binding energy prediction; therefore, the comparison of Lead Finder's accuracy with that of similar programs was limited to only those programs that are capable of predicting binding energy. The data on the rmsd between predicted and experimental binding energies are available for Glide XP:[17] 1.70 kcal/mol for 136 well docked complexes and 2.3 kcal/mol for all 198 complexes; for AutoDock 3.05[4] − 2.18 kcal/mol, for AutoDock 4[5] − 2.62 kcal/mol; however, it must be noted that in case of AutoDock, the same set of complexes was used for both calibration and assessment of the quality of binding energy prediction. On the basis of the above tests, Lead Finder appears to outperform these programs in the accuracy of binding energy prediction.

An analysis of errors in binding energy prediction with Lead Finder suggests that overscoring errors (cases when the predicted dG value is more favorable than the experimental value) are characteristic for relatively poor binders, while underscoring − for potent binders as shown in Figure 1. Among the 41 structures whose predicted dG value was in excess of 2 kcal/mol less favorable than the experimental value, 24 complexes had experimental dG values between

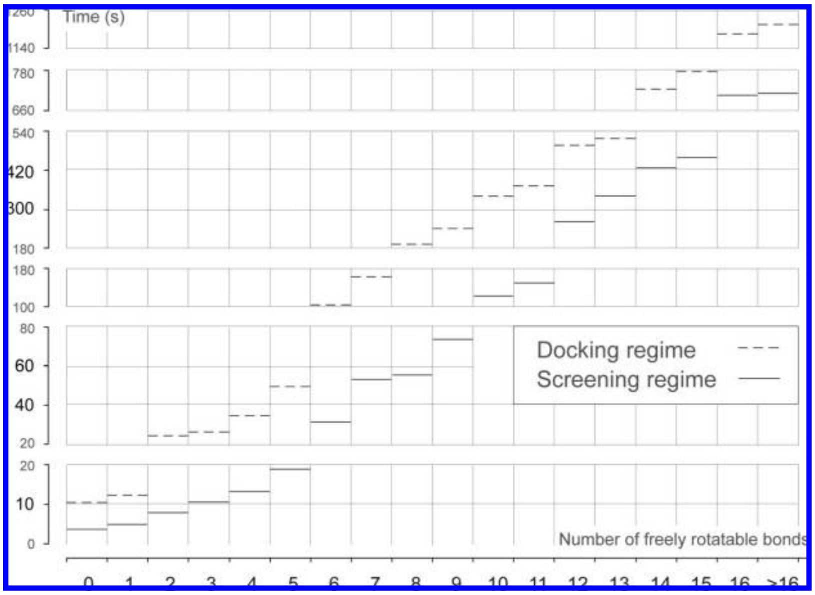

Figure 2. Dependence of the speed of docking calculations on the number of ligand's freely rotatable bonds in Lead Finder's docking and screening regimes. Each mark represents an average processing time for compounds from the docking test set of 407 complexes.

−8 and −12 kcal/mol, 15 − between −12 and −16 kcal/mol, and 2 structures below −16 kcal/mol. Only 6 structures revealed underscoring errors ranging from 3.0 to 3.8 kcal/mol, and for all other cases the error was less than 3 kcal/mol. As for the overscoring errors, 4 out of the 27 protein−ligand complexes revealed extremely poor binding (dG higher than −4 kcal/mol), 12 − low-specific binding (dG between −4 and −8 kcal/mol), and 11 were quite specific binders (dG lower than −8 kcal/mol). Only 7 overscoring errors were in excess of 3 kcal/mol, out of which 3 structures (1bma, 1tnk, 1tnl) corresponded to quite unspecific ligand binding (with dG above −8 kcal/mol). Such a trend indicates that for low-specific binders, for example 1tnk and 1tnl with experimental dG above −3 kcal/mol, our dG scoring function counts van der Waals interactions and nonpolar solvation in a continuous manner, even in the absence of tight protein−ligand contacts, leading to overscoring. The detailed information on the protein−ligand complexes comprising training and test sets, their calculated and experimental binding energies, and the physicochemical properties of ligands can be found in the Supporting Information.

2. Docking Success Rate. Our benchmarking study of Lead Finder docking success rate was set up in the spirit of recently proposed recommendations for the evaluation of computational methods.[21] In particular, all protein structures were prepared automatically as described above and in the Supporting Information; no optimization of protein heavy atoms was performed; all computational experiments were performed in multiple runs, and finally all relevant data, such as the protein and ligand structures and docking settings, were made publicly available for independent review by third parties.

The docking success rate obtained with Lead Finder on different test sets of protein−ligand complexes ranged from 80.0% (on GlideXP and FlexX test sets) to 96.0% (on Surflex and MolDock test sets) as shown in Table 3. These data can be used for an objective comparison with competitive programs since they have been benchmarked against the same sets of protein−ligand complexes according to the original publications.

Table 3 provides a summary of docking success rates for different programs and shows that Lead Finder has outperformed all docking programs for which the reliable original benchmarks exist. Specifically, Lead Finder successfully docks 73 structures (out of 200) which could not be correctly docked by FlexX, 52 structures (out of 282) for Glide SP, 52 structures (out of 268) for Glide XP, 30 structures (out of 134) for Gold on the original Gold test set, 13 structures (out of 85) for Gold on Astex diversity test set, 9 structures (out of 77) for MolDock, and 21 structures (out of 81) for Surflex. Unfortunately, an intersection of all test sets cited in this work results in null; therefore, it is impossible to analyze docking failures common to all the programs we examined. However, if we do not consider the recently developed Astex diversity set that has no intersection with any other set cited herein, we observe that the intersection of FlexX, Gold, Glide SP, Glide XP, MolDock, and Surflex test sets contains 70 structures and there is one structure, 1lic, that none of the above programs was able to dock correctly. 1lic represents adipocyte lipid-binding protein complexed with hexadecanesulfonic acid resolved to 1.6 Å. It is likely that the large number of freely rotatable bonds is the major reason for the incorrect ligand (hexadecanesulfonic) docking as it makes the search of a global minimum difficult because of the large number of possible ligand conformations. Interestingly, Lead Finder was able to successfully dock the ligand to the 1lic protein structure in 12 tests out of 20 with the mean rmsd of 1.69 Å obtained from the successful docking runs. It should be noted that correct solutions (within 2 Å from the PDB data) were also found in the unsuccessful docking runs on 1lic, but they had been slightly outscored by incorrect solutions before the final optimization took place.

As shown in Table 3, the docking success rate achieved by Lead Finder in the screening (faster, less accurate) regime was also found to be better than that of the majority of the competitive programs examined, with MolDock being the only exception. The attractiveness of the screening regime for high-performance calculations is illustrated in Figure 2; on average, it is 2−4 times faster than the docking regime. With the average time of 30 s needed to dock a ligand with up to 8 rotatable bonds on a modern CPU such as Intel Core2Duo 2.4 GHz under Linux, a small computer cluster of 32 CPUs can process about 100,000 ligands per day. We must also note that Lead Finder's screening and docking regimes revealed a striking similarity in binding energy predictions. That is, binding energies calculated for ligand poses obtained with either docking or screening regimes were quite close to each other. For the entire set of 407 protein−ligand complexes the rmsd of binding energies estimated in two distinct docking regimes was only 0.22 kcal/mol. This means the fast and less accurate screening regime was practically as effective as the more rigorous docking regime in the search of an energy minimum. This is good news for virtual screening applications where the geometry of ligand binding is usually not known in advance, but at the same time a reliable energy scoring is critically important for distinguishing between active and inactive compounds.

3. Docking Errors. Although Lead Finder has demonstrated a high overall docking success rate, it could not dock a number of ligands with 2 Å or better precision to their cognate protein structures. On the set of 407 protein−ligand complexes examined, 63 failures were observed in the docking regime and 92 - in the screening regime. Following the work of Kai Zhu et al.,[49] we considered two main sources of docking failures, namely − scoring and sampling errors. The first class of failures is related to the erroneous ranking of predicted ligand poses. Scoring errors are evident when an incorrect pose receives a better score than the correct one. The second class is generally related to the insufficient sampling of the ligand phase space, so that correct poses, irrelative of their rank, are not found during a docking run.

In the current study we take a closer look at scoring and sampling errors and introduce additional classification to segregate incorrectly docked complexes into groups for their further examination and analysis of the common causes of docking failures. For scoring errors, we classify cases when a top-ranked pose receives notably (over 1 kcal/mol) better energy than the correct pose as *evident* scoring errors. When this difference is less than 1 kcal/mol, we refer to such cases as *minor* scoring errors. There are also cases when the docking success criteria fail to identify satisfactory docking. For example, when the protein−ligand binding is quite weak, or some parts of the ligand do not form specific tight contacts with the protein, or the active site cavity is not deeply buried inside the protein but forms a shallow cleft on its surface, the rmsd between a predicted pose and the reference ligand pose may exceed 2 Å, and therefore it is considered a docking failure by the common standard of acceptance. Such cases are not rare for enzymes where natural substrates are abundant in the cell and binding constants are in submicromolar to millimolar range. In these situations, multiple substrate-ligand binding modes can be acceptable due to the relatively unspecific binding and/or pronounced degeneracy of the potential energy surface corresponding to the protein−ligand interactions. We classify these cases as *dubious* scoring errors. We leave the interpretation (admissibility or inadmissibility) of such errors to docking software users.

With sampling errors, cases when a reference ligand energy (binding energy estimated for the reference ligand) were notably better (over 1 kcal/mol) than the energy of a top-ranked pose and no poses close enough to the reference could be found, were classified as *evident* sampling errors. That means a docking run did not find the optimal solution. Another case is the *local* sampling error, in which poses substantially close (within 2 Å rmsd) to the reference ligand pose were found but not top-scored, presumably due to the inadequate local optimization. Another case of sampling error occurs when the root fragment of a relatively large ligand molecule is docked correctly, but its terminal branch is placed differently from the reference pose, resulting in the overall rmsd in excess of the 2 Å cutoff. This type of docking error is characteristic of molecules that have relatively long terminal branches with three or more sequentially joint fragments. We classify this type of docking error as the *terminal branch* error. A more exhaustive search of the phase space available for terminal branches is necessary in these cases.

Finally, we categorize dubious sampling errors as those that occur with big and complex active sites interacting with a ligand that has many conformational freedoms. The problem with dubious sampling errors is that frequently the reference pose has only a marginally better energy than other poses, and therefore a docking run may converge before the correct solution is found (or optimized).

According to the proposed classification, Tables 4 and 5 summarize the docking failures we observed with Lead Finder in the docking regime. Out of 63 docking failures we observed, 31 were attributed to scoring errors (Table 4), 33 - to sampling errors (Table 5), and 2 protein structures were considered irrelevant for docking studies due to severe inconsistencies in their active site geometry. Docking of 1mfe was found to possess both types of errors, a minor scoring error (the reference pose was slightly less favorable by its free energy than the top-scoring poses) and a sampling error (the reference pose was not consistently identified). As shown in Table 4, most of the evident scoring errors took place for relatively large ligands with 13 freely rotatable bonds or more. Two ligands (1bkm and 1n2v) were docked correctly in general (e.g., the ligand's pharmacophore group was correctly placed), while the failure of docking 5tim was accompanied by a quite unspecific binding of the corresponding ligand. Minor scoring errors accompanied the docking of 6 ligands, out of which 2 structures (1d3d and 1e5i) were docked correctly in general, and 2 structures (1sq5 and 1n2j) revealed relatively unspecific binding. Finally, 12 scoring failures were classified as dubious, out of which 10 ligands were docked correctly in general.

Out of the 33 cases of sampling errors presented in Table 5, 12 cases were classified as evident sampling errors. Of those 12 cases, 11 were attributed to the relatively large ligands with 15 freely rotatable bonds or more. Overall, 10 failures were assigned to an insufficient local optimization of top-ranked poses that were reasonably close to the corresponding reference pose; 4 of these cases revealed generally correct docking. Out of 6 cases with insufficient optimization of terminal fragments, 4 ligands were docked correctly in general. Finally, 5 scoring errors were recognized as dubious, and 3 of them were characterized by having a relatively unspecific ligand binding. For example, ligand binding sites in hemoglobin and calmodulin (structures 1g9v and 1ctr

Table 5. Docking Failures (In the Docking Regime) Attributed to Different Sampling Errors (E − Evident Error, L − Local, T − Terminal Branch, D - Dubious)[a]

PDB code	NFRB	rmsd, Å	dG ref, kcal/mol	error type
1adf	14	6.06	−12.3	T
1ake	20	2.62	−25.5	L
1apt	20	9.83	−12.0	E
1apu	17	8.94	−9.7	L
1apw	15	3.64	−11.1	E
1cnx	11	6.56	−9.5	T
1dih	15	3.13	−14.0	L
1eed	21	8.94	−11.5	EL
1glq	14	2.45	−11.3	L
1hef	21	3.77	−14.4	E
1hpx	15	2.62	−14.4	E
1hte	8	6.99	−8.6	L
1imb	8	2.07	−7.4	D
1ivd	3	4.75	−4.7	E
1jd0	1	4.34	−6.8	D
1meh	6	2.49	−8.3	T
1poc	24	2.70	−13.4	L
1ppi	22	4.99	−18.1	E
1ppm	17	4.51	−11.7	E
1tni	5	2.31	−6.5	D
2tsc	7	2.18	−12.8	T
4est	4	3.71	−10.4	T
4phv	15	4.02	−15.3	EL
4tmn	14	6.51	−13.0	E
5tmn	14	2.23	−12.8	T
6tmn	14	2.72	−11.3	E
9hvp	21	2.20	−15.3	L
1ctr	5	7.64	−7.2	D
1g9v	6	5.36	−7.2	D
1mfe	15	6.05	−7.8	E
3mth	3	4.20	−5.1	D

[a] The PDB codes that were docked correctly in general are highlighted in bold; the underlined PDB codes correspond to unspecific protein-ligand complexes. The binding energies (dG ref) for the optimized reference ligand pose are provided.

Table 6. Structures That Were Docked Incorrectly in the Screening Regime (E − Evident Sampling Error, L − Local Sampling Error)[a]

		dG top pose, kcal/mol		
PDB code	rmsd, Å	calc	ref	error type
1fh9	5.17	−9.2	−9.9	E
1s3v	7.83	−9.6	−9.6	E
7cpa	2.66	−14.4	−14.6	L
1rne	6.64	−15.8	−15.6	E
2upj	2.06	−10.9	−10.7	L
1f0u	3.18	−9.2	−10.1	L
1f0t	5.14	−9.1	−9.2	EL
1cdg	5.87	−7.9	−6.3	L
1ezq	8.87	−13.0	−13.4	E
121p	2.59	−16.2	−18.2	L
4fbp	2.74	−8.6	−9.1	L
1byb	3.02	−15.3	−16.9	L
1hps	3.77	−12.7	−13.2	EL
1cil	3.91	−8.6	−8.3	L
1dwc	4.09	−9.9	−10.1	E
1mzc	4.40	−9.9	−10.2	E
1ppk	4.61	−9.9	−10.0	E
1f0s	4.72	−9.9	−9.5	EL
1nco	7.25	−13.7	−14.2	L
1mcr	2.04	−6.3	−6.1	L
1f0r	2.28	−11.6	−11.7	L
3cpa	2.14	−9.2	−9.0	L
1pha	2.14	−9.5	−9.2	L
1ppl	2.16	−12.6	−13.1	L
1lpz	2.19	−12.6	−12.6	L
1aaq	2.27	−12.4	−13.2	L
1d0l	2.65	−10.5	−10.8	L
2ak3	2.84	−9.2	−9.2	L
1pph	2.90	−9.9	−9.9	L
1apv	2.97	−11.3	−11.7	L
1did	3.11	−7.2	−7.2	L
1jje	3.55	−11.7	−12.2	L
1lic	5.88	−7.7	−7.7	L
1uml	6.44	−10.4	−10.0	EL
7upj	2.15	2.1	−10.0	L

[a] The binding energies (dG) for the optimized reference ligand pose (ref) and the top-scored docked pose (calc) are provided.

correspondingly) are quite shallow in shape, so that roughly half of the ligand's surface comes into contact with the protein.

It should be mentioned that 2 docking failures (1gm8 and 1lmo) were not attributed to either scoring or sampling error because the corresponding protein structures were regarded as having inconsistencies that were incompatible with docking studies. For example, structure 1gm8 representing penicillin acylase complex with its natural substrate (penicillin G) actually corresponded to an inactive enzyme mutant (mutation N241A). This mutation disrupted the native binding mode of the substrate (which can be traced from structure 1gm9 and was studied in detail elsewhere[50]) and made it quite unspecific. Thus, we believe that structure 1gm8 was erroneously included in the Astex diverse set, which was aimed at selecting high-quality structures only; we suggest using structure 1gm9 instead (Lead Finder successfully deals with this structure). Structure 1lmo contains a chemically inconsistent ligand (with some of its single C−N bonds being only ∼1 Å in length), making ligand docking senseless without correcting the reference ligand structure and reoptimizing the protein−ligand complex.

4. Docking Errors in the Screening Mode. In total, 92 ligands could not be docked correctly in the screening regime (as compared to 63 in the docking regime), resulting in a docking success rate of 77.4%. Interestingly, 6 ligands (structures 1baf, 1g52, 1n2v, 1tni, 3mth, 9hvp) correctly docked in the screening regime but did not dock correctly in the docking regime. However, close examination of these cases reveals that due to less exhaustive sampling in the screening mode, a smaller number of poses was found with scores comparable to corresponding reference poses. As a result, there is a smaller chance for an incorrect pose to outscore the correct solution.

As shown in Table 6, almost all docking failures in the screening regime can be attributed to the sampling errors. To give a general illustration of the docking accuracy in the screening regime, we roughly classify sampling errors into the evident and local errors. Evident errors in the screening mode clearly point to cases of premature convergence (e.g., when the docking algorithm exits before the energy-optimal pose is found), while the more exhaustive docking regime proceeds to locate the global optimum. As shown in Table 6, the premature convergence takes place only in 11 out of 33 cases. An insufficient local sampling remains the major reason of docking failures in the screening regime.

5. Virtual Screening Performance. As mentioned in the section 'Overview of the Scoring Function', Lead Finder uses a special type of the scoring function called VS-score to rank ligands by their predicted affinity in docking-based virtual screening experiments. Although the energy components of

Table 7. Energy Scaling Coefficients and Their Ratio for the dG- and VS-Scoring Functions

	dG-score	VS-score	VS/dG
k_{VdW}	0.1344	0.10752	0.8
k_{sol}	0.2262	0.27144	1.2
$k_{polar\text{-}P}$	−0.2496	−0.2496	1
$k_{polar\text{-}S}$	0.3024	0.3024	1
$k_{nonpolar\text{-}P}$	−0.4	−0.8	2
$k_{nonpolar\text{-}S}$	−0.0896	−0.2688	3
$k_{elec,0}$ (buried)	0.050247	0.040197	0.8
$k_{elec,1}$ (intermediate)	0.0585	0.09945	1.7
$k_{elec,2}$ (surface)	0.094192	0.12245	1.3
k_{lig_born}	0.008775	0.004388	0.5
$k_{HB,lig\text{-}pen}$	−0.0626	−0.0626	1
$k_{HB,prot\text{-}pen}$	0.0691	0.04146	0.6
$k_{HB\text{-}polar}$	0.74	1.11	1.5
$k_{HB\text{-}nonpolar}$	0.0654	0.0981	1.5
$k_{HB,corr}$	−0.6272	−2.5088	4
k_{Me}	1	2	2
k_{nb}	0.093093	0.102402	1.1
k_{1-4}	0.030778	0.030778	1
$k_{dihedral}$	0.448	0.224	0.5
k_{tors}	0.1176	0.1176	1

the VS-scoring function are identical to the dG-scoring function, the corresponding energy-scaling coefficients are different. The reason for introducing the additional scoring function was to achieve the maximum recognition capability between active and inactive ligands. Thus, the parametrization of the VS-score implied an adjustment of scaling coefficients to achieve maximum efficiency in virtual screening experiments on the training set of protein targets and the corresponding active ligands (a single set of decoy ligands was used in all experiments). ROC values were chosen as an integral quantitative indicator of screening efficiency during VS-score parametrization. Sixteen proteins were chosen to construct a virtual screening training set, while the remaining 34 proteins were used as a test set, on which the screening efficiency was independently benchmarked. The division of targets into training and test sets was aimed at preserving diversity in protein functionalities in both sets and keeping the training set to a reasonable size.

The set of energy scaling coefficients obtained for the VS-scoring function is presented in Table 7, where comparison to the corresponding coefficients of the dG-scoring function is also provided. The following remarks outlining the distinct functionalities of Lead Finder VS- and dG-scoring functions can be made. First, the balance between van der Waals and solvation contributions is shifted toward solvation in the case of the VS-scoring function, which is obviously due to a more specific character of the latter type of interactions. Moreover, within the solvation contribution the surface-based terms take over the volume-based summand, assumingly due to a more specific (however, more computationally demanding) representation. The performance of Lead Finder in virtual screening appeared to be sensitive to this difference between the dG- and VS-scoring functions. Second, strengthening of energy contribution from H-bonds was found to be more important for the VS- than for the dG-scoring function. Again, that was in line with the more demanding character of VS-scoring function toward specific (e.g., having sharp distance and angular dependence) interactions. The enhancement of correlated hydrogen bonds was found to be especially useful. The contribution from electrostatic interactions, on the contrary, did not reveal steep influence on the virtual screening performance. The enhancement of contribution described with the intermediate electrostatic screening function (D_1) was found to be useful to avoid erroneous trapping of ligand's charges in uncharged protein cavities, but the benefit was not significant thereby suggesting that further improvement of the implicit electrostatic model is required. Finally, our experience suggests that the gap between the currently distinct dG- and VS-scoring functions is not significant. In fact, this makes sense from the theoretical standpoint.

As can be seen from Table 8, which illustrates Lead Finder's performance in the virtual screening experiments, the average ROC value for the training set is just slightly better than the one for the test set (0.93 vs 0.90). However, this difference stems mainly from 2 particular targets from the test set − thymidylate synthase (TS) and progesterone receptor (PR). For those two targets, a relatively poor enrichment (compared to other targets) was achieved.

The most outstanding enrichment (ROC = 1.00) was obtained for the oligopeptide-binding protein (OppA), for which a visual inspection of predicted poses of active ligands revealed correct docking; multiple hydrogen bonds and complementary electrostatic interactions of charged groups yielded high VS-scores of active OppA ligands. The same factors (multiple specific hydrogen bonds and electrostatic interactions) resulted in a high enrichment for orotidine-5′-phosphate decarboxylase (OPD) (ROC = 0.99). One notable achievement among the successful targets from the training set was that the peroxisome proliferator activated receptor gamma (PPAR-g), for which ROC = 0.98 was obtained with Lead Finder. However, this target had traditionally been recognized as difficult for virtual screening studies.[17,24,26] A visual inspection of predicted poses of PPAR-g active ligands revealed their correct docking (as can be judged from numerous PDB structures for PPAR-g and its inhibitors), with precisely mapped crucial hydrogen bonds and hydrophobic interactions. Moreover, we found that one protein model was sufficient for correct docking of all PPAR-g active ligands, though in some previous studies two models of protein structure were used[17] differing in the side chain position of phenylalanine 282. However, we found it reasonable to use two protein models (PDB structures 1e66 and 1eve) for acetylcholine esterase, due to the overlapping of some active ligands with phenylalanine 330 that can take two distinct conformations.

Although OppA, OPD, and PPAR-g (proteins from the training set) had no analogues among the 34 proteins studied in the virtual screening experiments, a number of other targets from the test set also revealed perfect enrichment. Active ligands of beta-secretase (ROC = 0.98) were correctly docked due to the location of crucial hydrogen bonds with catalytic aspartate residues and the positioning of the ligand's hydrophobic moiety inside the well-shaped protein's hydrophobic pocket. A successful enrichment for HIV-1 protease (HIVP) is also explained by the correct docking of active ligands and forming highly specific hydrogen bonds and hydrophobic contacts with the enzyme.

However, there were a number of targets for which Lead Finder failed to yield high enrichments. Glucocorticoid receptor (GR) revealed the worst results among the training set structures (ROC = 0.82). A similar situation with PR

Table 8. Results of the Virtual Screening Studies for 34 Protein Targets

target	PDB code	ROC	95% CI	EF20	EF40	EF70	no. of ligands	<dG>, kcal/mol
Test Set Proteins								
beta-secretase	1m4h	0.98	0.002	15.2	16.9	16.3	40	−10.8
HIV-1 protease	1pro	0.98	0.001	16.5	13.2	13.4	50	−11.5
factor Xa	1fjs	0.98	0.002	14.4	12.8	11.4	50	−9.7
estrogen receptor antagonists	3ert	0.97	0.001	37.73	23.83	15.24	30	−11.6
ribonuclease A	1qhc	0.95	0.005	45.3	12.9	8.9	30	−9
epidermal growth factor receptor kinase	1m17	0.95	0.002	5.4	7.3	8.1	50	−9.4
cAMP-dependent protein kinase	1fmo	0.94	0.003	7.9	6.2	6.6	50	−10.3
urokinase-type plasminogen activator	1gj7	0.94	0.002	5.8	6.9	7.3	20	−9.2
p38 MAP kinase	1kv2	0.92	0.002	3.1	4.2	5.4	50	−10.8
acetylcholinesterase	1e66/1eve	0.91	0.004	3.8	4.3	5.1	30	−8.2
HSP90	1uy6	0.89	0.003	2.9	3.9	4.5	30	−8.6
Lck kinase	1qpe	0.87	0.007	5.4	4.6	3.8	40	−8.3
estrogen receptor agonists	1l2i	0.86	0.004	1.65	2.29	2.66	30	−9.3
vascular endothelial growth factor receptor kinase 2	2oh4	0.86	0.006	7.2	4	3.7	50	−8.9
thermolysin	4tmn	0.86	0.012	22.4	16.6	3.7	20	−9.5
neuraminidase	2qwg	0.84	0.005	8.4	3.7	2.6	30	−7.3
thymidylate synthase	1f4g	0.77	0.012	4.5	3.2	2.3	15	−8.7
progesteron receptor	1sr7	0.76	0.013	2.7	2.1	2	20	−10.4
Training Set Proteins								
oligopeptide-binding protein	1b5j	1	0.001	111.8	89.4	78.3	16	−15
orotidine-5′-P decarboxylase	1eix	0.99	0.004	56	64	26.1	18	−11
protein tyrosine phosphatase 1B	1c84	0.99	0.002	73.8	55.4	11.7	20	−9.5
peroxisome proliferator activated receptor gamma	1fm9	0.98	0.001	9.6	11.8	11.7	50	−11.9
ribonuclease T1	1rnt	0.97	0.007	74.1	74.1	35.4	10	−8.5
thrombin	1c4v	0.96	0.002	12	8.6	10.8	40	−10.2
tyrosine kinase C-SRC	2src	0.96	0.002	11	12.1	7.7	50	−9.8
Trypsin	1qbo	0.95	0.003	9	9.4	9.5	20	−10.6
thymidine kinase	1kim	0.94	0.010	31.8	27.8	20.5	10	−8.9
mineralocorticoid receptor	2aa2	0.94	0.007	10.1	5.8	10.4	10	−11.2
poly(ADP-ribose) polymerase	1efy	0.92	0.004	3.5	5.7	7.6	10	−7.5
penicillopepsin	1bxo	0.91	0.010	6	8.2	5.5	6	−10.3
cyclooxygenase-2	1cx2	0.91	0.001	2.8	4	5.1	50	−11.3
fibroblast growth factor receptor kinase	1fgi	0.86	0.004	2.7	3	3.6	50	−9.6
angiotensin-converting enzyme	1o86	0.83	0.011	3	2.7	3.8	20	−9.4
glucocorticoid receptor	3bqd	0.82	0.007	1.7	2.5	2.7	50	−10.3

was observed for the test set targets (ROC = 0.76), while other nuclear receptors (including mineralocorticoid (MR) and estrogen receptors (ER_{ant} and ER_{ag})) demonstrated moderate to high enrichments. A close analysis of GR and PR reveals that their spacious ligand binding cavities, deeply buried inside the protein globule, can easily accommodate ligands of diverse shape and size; for this reason, big and hydrophobic ligands gain higher scores that outweigh the energy terms (such as hydrogen bonding energy and penalties) that provide for specific protein−ligand recognition. The local protein flexibility can also contribute to specific tuning of the receptor's binding pocket, so that native PR and GR ligands could be optimally positioned. This suggestion has been partially confirmed by our observation that GR structure 1m2z (used in the virtual screening study[26]) could not accommodate fairly large native GR binders; structure 3bqd performed better in that respect, though there were still problems with big active ligands. Finally, the Surflex test set of decoy ligands (as it was downloaded from the developer's site[42]) contained a number of active ligands for nuclear receptors; for technical reasons, those ligands were not filtered out from the decoy set, and a visual inspection of top-ranked ligands obtained in the virtual screening experiments with Lead Finder always revealed the active compounds that were initially claimed to be decoys.

Thymidylate synthase (test set target) also revealed a relatively poor enrichment (ROC = 0.77). In this case, we attribute results to the insufficiently specific binding of its native ligands (the predicted average binding energy of active ligands was only −8.7 kcal/mol), which is likely due to the relatively shallow and surface-exposed structure of the binding site of the enzyme. Such cases, as already discussed above, represent a significant scientific challenge for correct accounting of the surface-exposed hydrogen bonding and electrostatic interactions. It is probably that a similar situation takes place with neuraminidase (ROC = 0.84), where the binding site represents an extended funnel - ligands have too many degrees of freedom to slide along protein's surface, which dilutes the specificity of binding. As in the case involving TS, the native ligands of neuraminidase reveal a fairly modest binding energy (−7.3 on average). A moderate enrichment was obtained for angiotensin-converting enzyme as well (ROC = 0.83), and again we relate this observation to the possibility of big decoy molecules receiving high scores due to the spacious active site funnel of the enzyme. Despite the fact that all active ligands were coordinated with Zn^{2+} ion of the enzyme active site, this energy component was insufficient to outscore the big decoy ligands (many of which were also correctly coordinated to Zn^{2+} ion!).

Nonetheless, potent binding of native ligands is not the necessary condition for obtaining high enrichments, and the example of poly(ADP-ribose) polymerase (PARP) proves that. PARP active ligands have moderate binding energy as shown in Table 8, but due to a number of specific protein–ligand interactions (correlated hydrogen bonds) which take place for all PARP binders, it was possible to optimize the VS-score to encourage such interactions (and/or penalize their absence). This example clearly illustrates the difference between VS- and dG-scoring functions.

Finally, special attention has to be paid to protein kinases, which are now intensively studied as drug targets.[51] In this study, 7 protein kinases were assessed by Lead Finder in virtual screening experiments. Two of them were included in the training set and the remaining 5 – in the test set (Table 8). Overall, kinases revealed a good enrichment with ROC = 0.86 for fibroblast growth factor receptor kinase (FGFR) and vascular endothelial growth factor receptor kinase 2 (VEGFR) to ROC = 0.95 for epidermal growth factor receptor kinase (EGFR) and ROC = 0.96 for tyrosine kinase C-SRC (SRC). A close examination of predicted poses for active kinase ligands revealed a common feature determining the degree of virtual screening success. The targets where active ligands succeeded in forming crucial correlated hydrogen bonds with the hinge fragment of a kinase demonstrated higher enrichment than targets where such hydrogen bonds were observed less frequently. During our virtual screening experiments we tested a number of protein structure models obtained from different PDB structures and found that even subtle structural differences influence the probability of finding the correct ligand positioning (with crucial hydrogen bonds) in the enzyme active site. It is likely that the main reason for that stems from the relative mobility of N- and C-terminal lobes of a catalytic kinase domain: subtle changes of their relative positions influence the accessibility of the correct ligand-binding pose. This means that successful virtual screening searches of protein kinase inhibitors must account for this protein mobility or at least use a number of protein structure models.

As for all Lead Finder benchmarks, detailed results of virtual screening experiments (structures of active ligands for each target, predicted dG-score and VS-score, and ligand's position obtained during screening) can be found in the Supporting Information.

CONCLUSIONS

We have presented a novel algorithm for ligand docking and high precision scoring functions that guide the search algorithm toward finding a correct ligand position and determining the free energy of binding. The exploration of novel features of genetic algorithms, the application of various multilevel local optimization procedures, and the optimization of computational expenses resulted in a substantial improvement in docking accuracy compared to the currently available methods, as was supported by the extensive docking success rate benchmarking studies. The inclusion of original energy terms in the Lead Finder scoring functions and the adjustment of individual sets of scaling coefficients for pose ranking during docking, binding energy estimations, and virtual screening also increased accuracy of predictions as was shown in the current benchmarking study.

Current results suggest that there is still much room for mastering fundamental mathematic and computational approaches for global optimization as well as the description of protein–ligand interactions and the contributions of protein ligand binding to the free energy. We hope work in these directions will continue. We also hope that our current results obtained with Lead Finder will encourage scientists involved in applied research to engage molecular docking methods in their arsenal.

An evaluation of Lead Finder as well as the detailed data on protein and ligand structures used in this study can be requested at http://www.biomoltech.com.

Abbreviations. SCP, screened Coulomb potential; DOF, degree of freedom; EF, enrichment factor; ROC, receiver operator curve.

ACKNOWLEDGMENT

This work was supported in part by The Foundation for Assistance to Small Innovative Enterprises of the Russian Federation (contract 4935r).

Supporting Information Available: Detailed description of results of docking success rate, binding energy estimations, and virtual screening benchmarking studies. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.

(2) Totrov, M.; Abagyan, R. Flexible protein–ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 215–220.

(3) Tietze, S.; Apostolakis, J. GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1657–1672.

(4) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(5) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–1152.

(6) 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.

(7) Thomsen, R.; Christensen, M. H. MolDock: A new technique for high-accuracy molecular docking. *J. Med. Chem.* **2006**, *49*, 3315–3321.

(8) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided. Mol. Des.* **1995**, *9*, 113–130.

(9) Rarey, M.; Kramer, B.; Lengauer, T. Multiple automatic base selection: Protein–ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369–384.

(10) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX Incremental Construction Algorithm for Protein–Ligand Docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228–241.

(11) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(12) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. eHiTS: an innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006**, *7*, 421–435.

(13) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198–212.

(14) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 333–344.

(15) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.

(16) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Glide: A New Approach for Rapid, Accurate Docking and Scoring.

1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(17) 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. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(18) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing Protein−Ligand Docking Programs Is Difficult. *Proteins: Struct., Funct., Bioinformat.* **2005**, *60*, 325–332.

(19) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A. New Test Set for Validating Predictions of Protein−Ligand Interaction. *Proteins: Struct., Funct., Genet* **2002**, *49*, 457–471.

(20) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(21) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 133–139.

(22) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct., Funct., Bioinformat.* **2004**, *57*, 225–242.

(23) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.

(24) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(25) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for Docking. *J. Med. Chem.* **2005**, *48*, 3714–3728.

(26) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(27) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902–1921.

(28) Mehler, E. L. Self-Consistent, Free Energy Based Approximation To Calculate pH Dependent Electrostatic Effects in Proteins. *J. Phys. Chem.* **1996**, *100*, 16006–16018.

(29) Mehler, E. L.; Guarnieri, F. A Self-Consistent, Microenvironment Modulated Screened Coulomb Potential Approximation to Calculate pH-Dependent Electrostatic Effects in Proteins. *Biophys. J.* **1999**, *75*, 3–22.

(30) Gasteiger, J. Empirical approaches to the calculation of properties. In *Chemoinformatics: a textbook*, 1st ed.; Gasteiger, J., Engel, T. Eds.; Wiley-VCH: Darmstadt, Germany, 2003; Vol. 1, pp 329−337.

(31) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators. *J. Comput. Chem.* **2002**, *23*, 517–529.

(32) Stouten, P. F. W.; Frommel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simul.* **1993**, *10*, 97–120.

(33) Payne, A. W. R.; Glen, R. C. Molecular recognition using a binary genetic search algorithm. *J. Mol. Graph.* **1993**, *11*, 74–91.

(34) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein−ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–D526.

(35) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(36) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(37) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2006**, *00*, D1−D4.

(38) The original Gold test set. http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/original_gold_test_set/ (accessed Sept 2, 2008).

(39) CCDC/Astex test set. http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/astex/pdb_entries/ (accessed Sept 2, 2008).

(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(41) Zhang, J.-W.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano, T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* **2004**, *28*, 401–407.

(42) Surflex test sets. http://www.jainlab.org/downloads.html (accessed Sept 2, 2008).

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

(44) Unpublished results.

(45) Data sets of the protein and ligand complexes used in the current work are available for download. http://www.biomoltech.com/downloads/target_proteins.zip, http://www.biomoltech.com/downloads/active_ligands.zip, http://www.biomoltech.com/downloads/surflex_decoy_ligand_set.zip (accessed Sept 2, 2008).

(46) Jain, A. N. Bias, reporting and sharing: computational evaluation of docking methods. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 201–212.

(47) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(48) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890.

(49) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. Long Loop Prediction Using the Protein Local Optimization Program. *Proteins: Struct., Funct., Bioinformat.* **2006**, *65*, 438–452.

(50) Chilov, G. G.; Stroganov, O. V.; Svedas, V. K. Molecular modeling studies of substrate binding by penicillin acylase. *Biochemistry (Moscow)* **2008**, *73*, 56–64.

(51) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there. *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

CI800166P