# Ligand-Based Structural Hypotheses for Virtual Screening

Ajay N. Jain*

*UCSF Cancer Research Institute and Comprehensive Cancer Center, University of California, San Francisco, California 94143-0128*

*Received October 13, 2003*

The majority of drug targets for small molecule therapeutics are proteins whose three-dimensional structure is not known to sufficient resolution to permit structure-based design. All three-dimensional QSAR approaches have a requirement for some hypothesis of ligand conformation and alignment, and predictions of molecular activity critically depend on this ligand-based binding site hypothesis. The molecular similarity function used in the Surflex docking system, coupled with quantitative pressure to minimize overall molecular volume, forms an effective objective function for generating hypotheses of bioactive conformations of sets of small molecules binding to their cognate proteins. Results are presented, assessing utility of the method for ligands of the serotonin, histamine, muscarinic, and $GABA_A$ receptors. The Surflex similarity module (Surflex-Sim) was able, in each case, to distinguish true ligands from random compounds using models constructed from just two or three known ligands. True positive rates of 60% were achieved with false positive rates of 0−3%; the theoretical enrichment rates were over 150-fold compared with random screening. The methods are practically applicable for rational design of ligands and for high-throughput virtual screening and offer competitive performance to many structure-based docking algorithms.

## Introduction

Discovery of novel lead compounds through computational exploitation of experimentally determined protein structures, either derived from screening of databases or through focused design exercises, is well established.[1] However, it is more frequently the case that a discovery effort lacks a high-resolution structure of the target protein. For example, membrane spanning G-protein-coupled receptors (GPCRs) or ion channels were the targets for nine of the top 20 selling prescription drugs worldwide in the year 2000,[2] and it is unlikely that general methods for solution of structures of these protein classes will be developed in the short term. Other classes of targets for which structure-based approaches will be challenging, such as membrane transporters, also form important targets for modeling. It is no less important in these cases to have predictive models of ligand activity than in the cases for which protein structure is known.

In these cases, predictive models of ligand binding to target active sites can be useful in at least two situations: (1) where many existing ligands are known but where they share side-effects or biological properties that limit their biological utility, and (2) where a small number of ligands have been discovered for a target (e.g. by high-throughput screening) that has not been extensively probed and augmentation of the set is a primary goal of a medicinal chemistry effort. In either case, improvements in potency, selectivity, or ADME characteristics may be required. Such improvements often require different chemical scaffolds, so computational approaches that are not strongly scaffold-dependent are preferable. The fundamental problem is to generate a hypothesis for how the ligands must be binding to their target, which then can be exploited through computational means. The focus of the work reported here is on the automatic construction of such models based on small numbers of competitive noncovalently binding ligands. The goals is to induce models that are specific enough to distinguish true competitive ligands from large pools of nonbinding ligands. Precise quantitative prediction of binding affinities is not the focus of this effort, since a number of approaches to that problem exist. But those approaches, whether they are grid-based as in the seminal work of Cramer et al.[3], multipoint pharmacophore-based,[4−6] or surface-based,[7−9] all depend on some hypothesis of the joint molecular poses of input molecules.

Given a small number of potentially quite flexible molecules of diverse chemical structures, one must generate a hypothesis consisting of a single pose for each input molecule such that the joint superposition of all molecules will lead to predictive models of biological activity. Instead of quantifying performance using sets of ligands whose protein-bound structure was known, the approach here was to take four therapeutically interesting targets with unknown three-dimensional structure (the muscarinic, serotonin, histamine, and $GABA_A$ receptors) and generate molecular superpositions using two or three known ligands for each. Evaluation of correctness follows recent molecular docking literature in quantitatively assessing theoretical screening utility, expressed as the degree to which ligands of the respective receptors can be distinguished from random compounds.[10−13] There are a large number of methods for computing molecular similarity and for producing molecular superpositions (Lemmen and Len-

* To whom correspondence should be addressed. Dr. Ajay N. Jain, UCSF Cancer Center, 2340 Sutter Street, #S336, Box 0128, San Francisco, CA 94143-0128. Phone: (415) 502-7242. Fax: (650) 240−1781. E-mail: ajain@cc.ucsf.edu.

10.1021/jm030520f CCC: $27.50 
Published on Web 01/17/2004

gauer have published an excellent review[14]), but most of these methods have not been evaluated for screening utility.

The methods for molecular similarity and alignment optimization used in the Surflex docking system[10,15] were coupled with a functional term to minimize the volume of molecular superpositions to construct an objective function for scoring superpositions of multiple molecules. Molecular superpositions that maximized the objective function were used as targets for virtual screening. Performance was comparable to many docking methods, with true positive rates of 60% at false positive rates of 0−3%. Many ligands of the target receptors with widely differing underlying chemical scaffolds were detectable at false positive rates of less than 1%.

Surflex's similarity module (referred to as Surflex-Sim in what follows), which implements the algorithms described here, is available free of charge to academic researchers for noncommercial use (see http://jainlab.ucsf.edu for details on obtaining the software). Molecular data sets presented herein are also available.

## Methods

The methodology reported here takes a small number of input molecules and yields a superposition of the molecules that optimizes an objective function that is composed of a molecular similarity component,[15] a restraint against excessive volume that enforces parsimony, and a restraint against ligand self-clashing (inappropriate atomic overlap between nonbonded atoms). In assessing performance of such methods, there are two main criteria: (1) assessment of the degree to which the predicted overlap matches with experimental observation; and (2) assessment of the utility of the model comprised by the overlap in identifying novel ligands with high sensitivity and specificity.

Since the focus of this paper is on applications to proteins whose structure is not known, the emphasis is on the second criterion. However, since benchmarks for screening utility exist for structure-based docking methods,[10,11] these are also used for direct comparison. Note that there is an intrinsic difficulty in assessing ligand-based models and in comparing them with protein structure-based methods. Since models are based on ligand structures, there is an inherent inductive bias in their construction that favors molecules that are highly similar to those molecules used for model construction. Consequently, molecular series should be employed which allow for extrapolation beyond the underlying chemical scaffolding used in the input molecules.

**GPCR and $GABA_A$ Data Sets.** Serotonin, muscarinic, histamine, and $GABA_A$ ligands (shown in Figure 1) were used to generate optimal molecular superpositions, which were then used to score other known ligands of the receptors relative to a background of random molecules. The two serotonin ligands (molecules A and B) were initially reported by Lin et al.[16,17] and were the subject of an extensive 3D QSAR study using Compass.[9] These ligands were chosen since they represent a snapshot in time of a particular group's medicinal chemistry efforts, which resulted in a molecular series that is quite different than many other

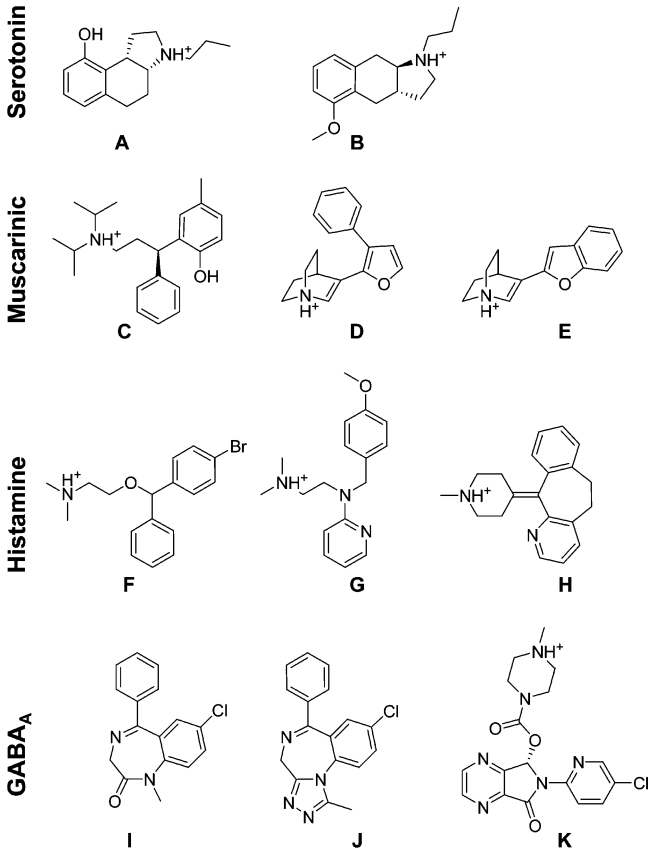


**Figure 1.** Molecules used for generation of GPCR and benzodiazepine binding site structural hypotheses. Molecules A and B are 5-$HT1_A$ ligands from a family of linear and angular tricyclic compounds that include both agonists and antagonists of both the serotonin and dopamine receptors (A has 0.1 nM binding affinity for 5-$HT1_A$, B 17 nM). Molecules C (tolterodine), D, and E are muscarinic antagonists (respective affinities of 0.3 nM, 9 nM, and 59 nM; functional inhibition of guinea pig urinary bladder contraction of 14 nM, 3 nM, and 33 nM). Molecules F−H (bromodiphenhydramine, pyrilamine, and azatadine) are H1 receptor antagonists (respective affinities of 13 nM, 2 nM, and 11 nM). Molecules I−K are $GABA_A$ receptor agonists (diazepam, alprazolam, and zopiclone) with nanomolar binding affinities.

ligands of the same cognate receptor. Binding affinities for the series were determined by radioligand displacement and were 0.1 nM for A and 17 nM for B. Since the tertiary amine nitrogen of each molecule is expected to be protonated at physiological pH, but the proton can be oriented in two ways, there are several potential superpositions of the molecules, despite their relative rigidity. There is a significant difference in their binding affinities, but they are known to be competitive for the same binding site. The three muscarinic antagonists (molecules C−E) include tolterodine, an approved therapeutic for urge-based urinary incontinence, and two quinuclidinene-based ligands developed by the same company.[18−20] Again, these represent a very small space of muscarinic antagonist scaffolds, and they also represent a single company's efforts. By contrast, the histamine antagonists (molecules F−H: bromodiphenhydramine, pyrilamine, and azatadine) represent first-generation antihistamines (respective US patent years: 1950, 1950, and 1967).[21] Molecules I−K are $GABA_A$ receptor agonists, all of which bind at the benzodiazepine binding site. Molecules I and J are

**Table 1.** The 100 Positive Molecules Used to Test the Models, with Annotation of Known Targets. Note: bcce Is Ethyl $\beta$-Carboline-3-carboxylate

| molecule | Ser | Musc | Hist | $GABA_A$ | molecule | Ser | Musc | Hist | $GABA_A$ |
|---|---|---|---|---|---|---|---|---|---|
| abecarnil | | | | x | methscopolamine | | x | | |
| alosetron | x | | | | methysergide | x | | | |
| alpidem | | | | x | metitepine | x | | x | |
| amitriptyline | x | x | x | | mianserin | x | x | x | |
| amoxapine | | x | x | | midazolam | | | | x |
| astemizole | | | x | | molindone | | x | | |
| atropine | | x | | | nefazodone | x | | | |
| bcce | | | | x | nortriptyline | x | x | x | |
| benztropine | | x | x | | olanzapine | x | | | |
| bethanechol | | x | | | ondansetron | x | | | |
| bromolysergide | x | | | | oxetorone | x | | | |
| brompheniramine | | x | x | | oxybutynin | | x | | |
| bupropion | | | x | | perospirone | x | | | |
| carbachol | | x | | | perphenazine | | x | x | |
| carbinoxamine | | | x | | phenindamine | | | x | |
| cetirizine | | | x | | phenoxybenzamine | | | x | |
| chlorpheniramine | | x | x | | pilocarpine | | x | | |
| chlorpromazine | x | x | x | | pimozide | | x | | |
| clemastine | | | x | | pindolol | x | | | |
| clobazam | | | | x | pizotyline | x | | | |
| clomipramine | | x | x | | prazepam | | | | x |
| clozapine | x | x | x | | prochlorperazine | | x | x | |
| cocaine | | x | | | procyclidine | | x | | |
| cyproheptadine | | x | x | | promethazine | | x | x | |
| darifenacin | | x | | | protriptyline | | x | x | |
| desipramine | | x | x | | quazepam | | | | x |
| dicyclomine | | x | | | ramosteron | x | | | |
| dolasetron | x | | | | risperidone | x | | x | |
| dotarizine | x | | | | ritanserin | x | | x | |
| doxepin | | x | x | | sertindole | x | x | x | |
| estazolam | | | | x | suriclone | | | | x |
| fluphenazine | | x | x | | telenzipine | | x | | |
| flutoprazepam | | | | x | terfenadine | | | x | |
| granisteron | x | | | | tetrazepam | | | | x |
| halazepam | | | | x | thiethylperazine | | | x | |
| haloperidol | | x | | | thioridazine | | x | x | |
| hydroxyzine | | | x | | thiothixene | | x | x | |
| iloperidone | x | | | | tiotropium | | x | | |
| imipramine | | x | x | | trazodone | | | x | |
| itasetron | x | | | | triazolam | | | | x |
| ketanserin | x | | | | trifluoperazine | | x | x | |
| levocabastine | | | x | | triflupromazine | | x | x | |
| lidocaine | | x | | | trimipramine | | x | x | |
| loratadine | | | x | | tripelennamine | | | x | |
| loxapine | | x | x | | triprolidine | | | x | |
| maprotiline | | x | x | | tropisetron | x | | | |
| meclizine | | | x | | zaleplon | | | | x |
| mesoridazine | | x | x | | ziprasidone | x | | | |
| metergoline | x | | | | zolpidem | | | | x |
| methotrimeprazine | | | x | | zotepine | x | | | |

classically structured agonists, and K is the first of a family of nonbenzodiazepine agonists displaying similar pharmacology to the classic agonists.[21]

To test the models that resulted from these input structures, GPCRDB (www.gpcr.org)[22] was used to identify GPCR ligands with measured binding affinity to any of the receptors within a particular family of better than 500 $\mu$M (muscarinic receptor types M1−5, serotonin receptor types 5-HT1−7, and histamine receptor types H1−4). The Merck Index was used to identify benzodiazepine receptor agonists.[21] Reasonable effort was made to be complete, with the resulting name list being cross-referenced against the Current Medicinal Chemistry database to yield 85 GPCR ligands of widely varying chemical structure and 15 $GABA_A$ agonists, of which nine were variations on the classic benzodiazepine scaffold and six were of varying chemotypes. Table 1 lists the names and annotated target specificity for all 100 molecules. As a negative control for all cases, ACD screening molecules were used, following two recent docking benchmarks.[10,11]

All of the molecules above were subject to the same preparation procedures, which involved automatic protonation, ring search, protonated nitrogen inversion, and minimization using a Dreiding-type force-field. Redundant conformations (those that did not have different ring geometries) were eliminated. Of these, the lowest energy conformations were retained (up to a maximum of 10) for each molecule. Note that typically, larger and more complex ring systems resulted in more unique conformations than smaller ones. These starting conformations accounted for alternative ring conformations and protonation geometries, with the flexibility due to acyclic bonds to be addressed by the optimization algorithms discussed below. The ACD screening set[11] originally contained 990 molecules, and of these, 850 were correctly processed and used as a negative control. The search and minimization procedure had a signifi-

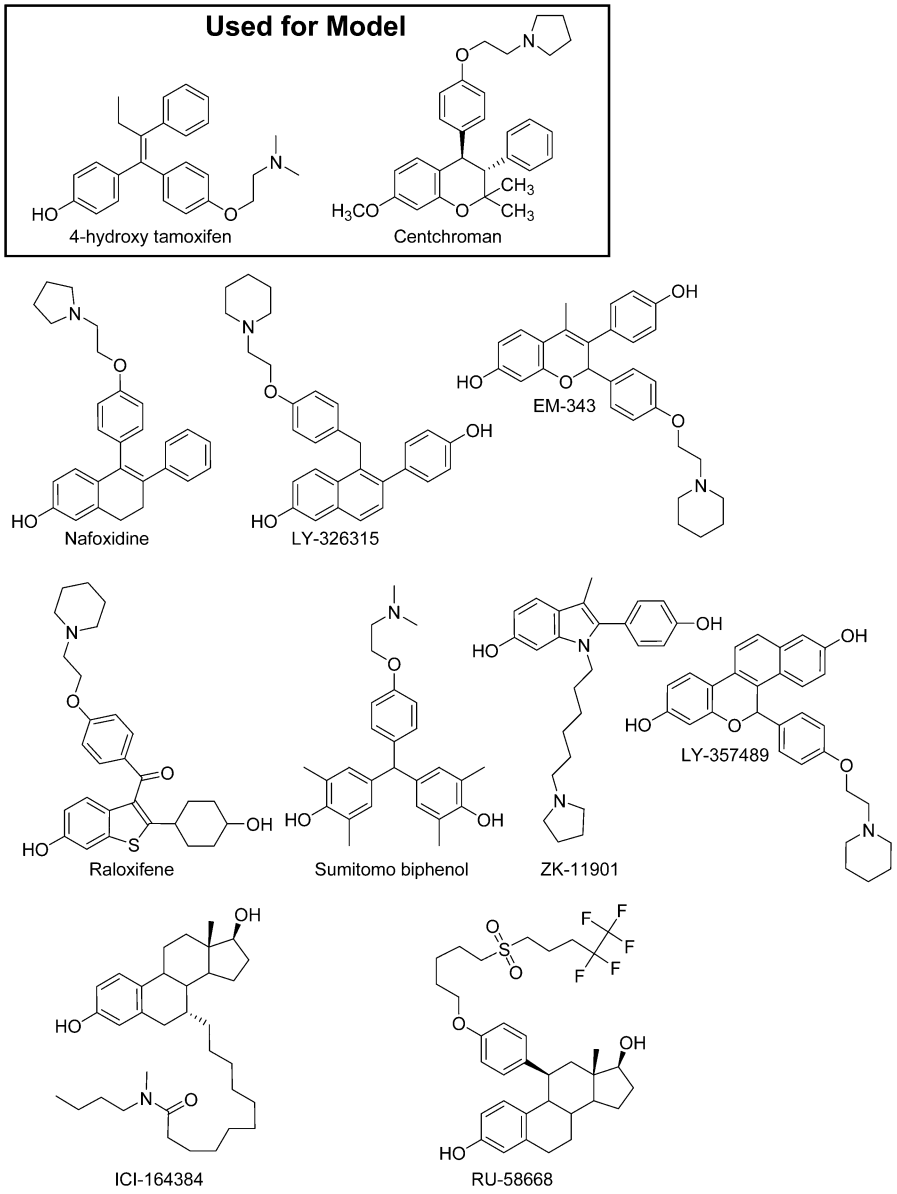


**Figure 2.** Estrogen receptor ligands used for protein-structure independent test of screening sensitivity. Centchroman and 4-hydroxytamoxifen were used to generate a ligand-based hypothesis. The remaining nine molecules are shown in the order in which they ranked in a screen including 941 random compounds. The ranks of the nine compounds were: 1, 2, 3, 4, 5, 6, 7, 62, 135.

cant impact on model generation, since molecules A, B, H, I, and J have flexible rings and the first three have asymmetric nitrogens that are subject to inversion. The total number of conformations was 78 for the 11 input molecules, 350 for the 85 novel GPCR ligands, 43 for the 15 $GABA_A$ agonists, and 1349 for the 850 screening molecules. It is important to stress that the molecular superposition methods described below sample the conformational space of the ligands much further than this initial sampling, but the on-line search is currently limited to acyclic bonds, necessitating this two-step approach.

**Docking Data Sets.** To provide a direct test of superposition accuracy and a direct comparison to the utility of structure-based screening methods, the estrogen receptor and thymidine kinase benchmarks from Bissantz et al.[11] were used. These consisted of 10 known ligands for each protein target and 990 compound ACD screening set mentioned above. The issues of inductive bias are present to a greater extent in these cases compared to the GPCR and $GABA_A$ sets, since the known ligands from the benchmarks have a relatively high degree of scaffold similarity. For the ER case, a very old nonsteroidal estrogen antagonist (centchroman) was used in conjunction with 4-hydroxytamoxifen to generate a molecular superposition. The additional ER ligand centchroman was subject to ring search and minimization as for the GPCR data sets, resulting in 10 initial conformations for model induction. For the TK case, deoxythymidine and acyclovir (the oldest among the purine nucleoside HSV TK inhibitors) were used. Figures 2 and 3 show the ligands for both cases and highlight the pairs chosen for model induction. To provide a direct comparison to the docking benchmark from which these examples were derived, the molecules were used unmodified from the original paper.[11] For computational efficiency, molecules with greater than 21 rotatable bonds in the screening set were excluded,

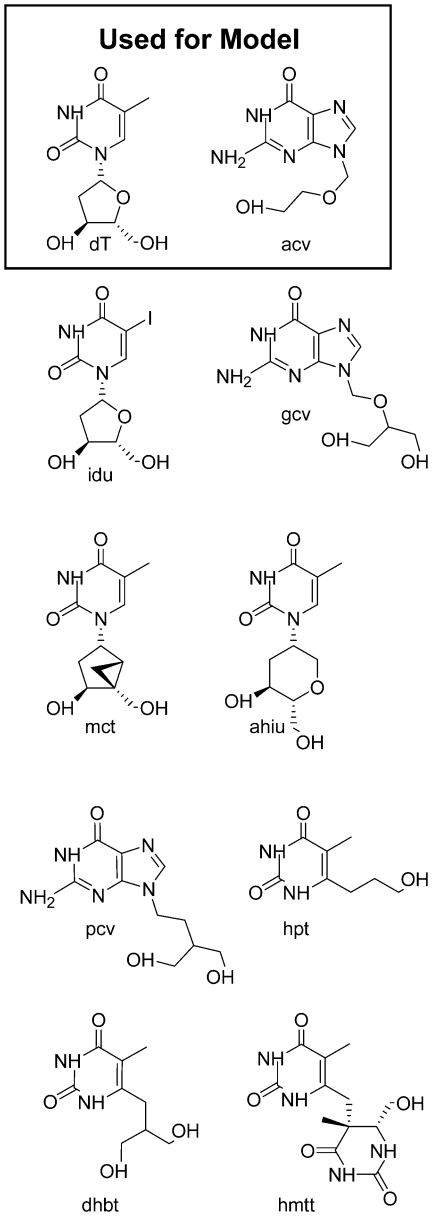


**Figure 3.** HSV-1 Thymidine kinase inhibitors used for protein-structure independent test of screening sensitivity. The abbreviations are as follows: dT, deoxythymidine; idu, 5-iododeoxyuridine; hpt, 6-(3-hydroxypropylthymine); ahiu, 5-iodoracil anhydrohexitol nucleoside; mct, (North)-methanocarbathymidine; hmtt, (6-[6-hydroxymethy-5-methyl-2,4-dioxo-hexahydro-pyrimidin-5-yl-methyl]-5-methyl-1*H*-pyrimidine-2,4-dione; dhbt, 6-(3-hydroxy-2-hydroxymethylpropyl)-5-methyl-1*H*-pyrimidine-2,4-dione; acv, aciclovir; gcv, ganciclovir; pcv, penciclovir. Deoxythymidine and acyclovir were used to generate a ligand-based hypothesis. The remaining eight molecules are shown in the order in which they ranked in a screen including 941 random compounds. The ranks of the eight compounds were: 1, 5, 6, 7, 8, 9, 10, 15.

resulting in 941 of the 990 molecules being used in the negative control experiments. This exclusion had negligible impact on the computed statistics. Note also that while 21 rotatable bonds is large with respect to druglike molecules, fewer than 100 of the molecules had more than 15 rotatable bonds.

**Computational Methods.** As discussed previously, the focus of this work is on the first step of the 3D QSAR problem: finding a superposition of active molecules that is sufficiently reflective of the true relative binding modes that construction of predictive models is possible. The method makes use of the morphological similarity approach[15] that is used in the Surflex molecular docking system.[10] The goal is to find a molecular superposition that maximizes the joint similarity of a set of N input molecules by varying their molecular poses, subject to the constraint that superpositions with smaller volumes are favored over those with nominally equivalent similarity but which occupy more space. The procedure is incremental in nature, building upon high-scoring pairwise superpositions one molecule at a time.

Procedures such as the MultiSEAL method of Feher and Schmidt[23] and the SQ method of Miller et al.[24] are similar in approach. The former method extends SEAL by using a greedy build-up to augment pairwise superpositions. The latter uses a genetic algorithm to search the space of fit molecular alignments. The Surflex-Sim methodology differs primarily in the choice of objective function and in that the alignment optimization relies less strongly on atom-based matching in favor of matching molecular surfaces, as described below.

**Molecular Similarity and Superposition.** Surflex-Sim utilizes the morphological similarity function and fast pose generation techniques described previously[15] to generate putative alignments of molecules or molecular fragments to other molecules. Briefly, morphological similarity is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid. The surface distances computed include both distances to the nearest atomic surface and distances to donor and acceptor surfaces. Since the function is dependent on the relative alignment of two molecules, the problem of optimizing the similarity of one molecule to the fixed conformation of another efficiently is a critical issue. Both the alignment and conformation of the molecule must be optimized.

The alignment problem can be addressed with an efficient algorithm because the molecular observations that underlie the similarity function are local and are not dependent on the absolute coordinate frame. So, two unaligned molecules or molecular fragments that have some degree of similarity will have some corresponding set of observers that are seeing the same things. Optimization of the similarity of two unaligned molecules is performed by finding sets of observers of each molecule that form triangles of the same size, where each pair of corresponding points in the triangles are observing similar features. The transformation that yields a superposition of the triangles will tend to yield high-scoring superpositions of the molecules. Figure 4 illustrates the similarity function and alignment optimization method. This function is able to separate pairs of molecules known to bind the same proteins from random pairs of molecules much better than methods based on 2D graph-theoretic measures of molecular similarity.[15]

The problem of flexibly aligning one molecule onto another is addressed as described previously[15] and is very briefly summarized below. The overall approach is a divide and conquer algorithm, making use of molecular fragmentation and incremental construction to ameliorate the exponential dependence of conformational space on number of rotatable bonds.

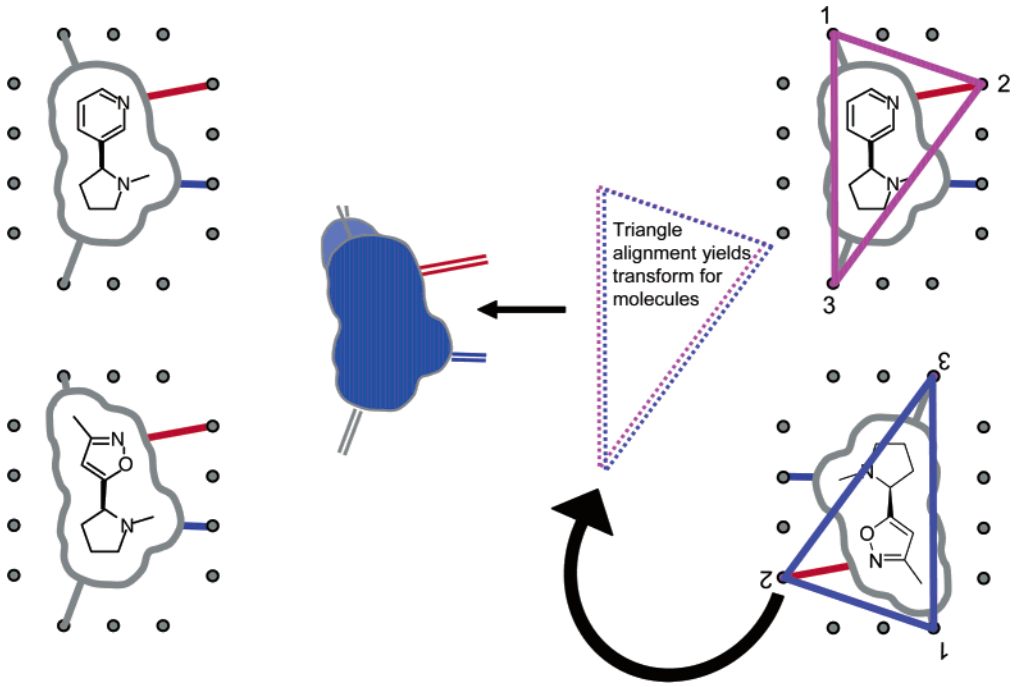


**Figure 4.** Left: Similarity between molecules is defined as a function of the differences in surface measurements from observation points. Ligands shown are nicotine and a competitive nicotinic agonist. While the pyridine and oxazole are very different based on a 2D graph-based assessment, they display very similar hydrophobic and polar surfaces. The similarity is a Gaussian function of the differences in the distances from the observation points to the surfaces. It is the difference in the distances to the blue surface shapes that drives the similarity metric. Right: The alignment optimization procedure seeks to identify corresponding triplets of observer points of the two molecules subject to two constraints: (1) the triangles must be of similar size, and (2) each of the points must be "seeing" similar molecular features. That is, an observer point of molecule 1 that is 2 Å from a hydrogen-bond acceptor oriented toward it must be corresponded with such an observer point of molecule 2. Triangle matches (such as the match between the blue and purple triangles) are accumulated and tested by applying the transformation to one molecule that superimposes the triangles of both.

**Molecular Fragmentation.** Molecular flexibility is addressed by molecular fragmentation. Molecules are fragmented by breaking acyclic rotatable bonds. Each such break eliminates a bond for conformational search and eliminates the need to cross the conformations of the two fragments. So, a molecule with seven rotatable bonds, where each bond is sampled at six rotameric positions, is reduced from $6^7$ (>250 000) conformations to $6^3 + 6^3$ (432) conformations, a reduction by nearly 3 orders of magnitude. In practice, a heuristic set of rules are employed in conformational sampling, where two, three, or six rotamers are used for each bond (e.g. 3 for $sp^3-sp^3$ bonds). Also, a maximum number of conformations per fragment can be specified (default 20), and the algorithm selects the most different conformations based on rmsd. Following completion of fragmentation and conformational search (and fast internal clash relaxation), the resulting molecular fragments are aligned to the target molecule. Those aligned fragments that score the best are retained and serve as input to the next step in the procedure. Note that the procedure of fragmentation, search, alignment, and scoring is completely automatic.

**Incremental Construction.** The highest scoring aligned fragments from above are used as "heads", from which a directed alignment of the "tail" (next molecular fragment) occurs by aligning each conformation of the appropriate fragment based on similarity to the target molecule, but subject to the constraint that the alignments generated must place the connector atom proximal to where it must be to make a connection to the head. As fragments are added to the growing partial molecules, similarity to the target molecule is optimized by gradient descent. In this process, a term to penalize internal clashes (taken directly from the Hammerhead scoring function) is employed. The process of incremental construction iterates on the best partial solutions until the molecule to be aligned has been fully reconstructed, resulting in fully optimized poses of the input molecule relative to the fixed conformation of the target.

**Superposition of Multiple Molecules.** The foregoing has addressed the superposition of one molecule onto a fixed pose of another. The joint superposition of multiple molecules is significantly more difficult than the superposition of one molecule onto a specific pose of another. The obvious difficulty is the combinatorial problem, since the number of configurations increases multiplicatively with additional molecules. For a molecule pair, given a sufficiently dense sampling of the conformations of the target molecule, the optimal mutual superposition of the two molecules is tractable.

Another difficulty is more subtle. Given four molecules, where each of two pairs are highly similar, there may exist a superposition of all four that yields very good matches within each similar pair, but which does not address the overlap between the two sets of pairs well. The obvious objective function for superimposing $N$ molecules is the sum of all pairwise molecular similarities,[15] but it suffers from this deficiency. An additional term, which seeks to overlay the notion of parsimony appears to alleviate this problem. Surflex-

Sim employs a joint similarity objective function that is the product of (1) the sum of all pairwise similarities and (2) the total empty volume in a sphere of fixed size centered on the superimposed ligands. This biases the solutions of joint superposition to the smallest possible volume, given equivalent joint similarities.

Using this function, Surflex-Sim employs a greedy strategy to identify superpositions of $N$ molecules that maximize the function. First, all pairwise superpositions are explored, beginning from a sampling of conformational space for each of the $N$ molecules. The best of these (100 by default) based on the objective function are retained. For each of these partial structural hypotheses, the remaining $N - 2$ molecules are aligned to each of the molecules within the superimposed pair, and the overall score is computed. The best hypotheses, which have grown by one molecule (each now containing three molecules), are retained, and the process repeats, with superimpositions increasing by one molecule per iteration, until all hypotheses have $N$ molecules.

**Overall Objective Function.** The overall objective function is the product of the mean pairwise similarity and the volume component, which is computed as follows. Given an alignment of $N$ molecules, an amount proportional to the excluded volume of a sphere of radius 10.0 Å is computed. This volume is approximated by uniform tesselation of the sphere by 66 points. The minimum distance from each point to *any* of the molecules in the alignment is computed (just as in the molecular similarity computations). The sum of all 66 such distances is computed and scaled down by a constant (528.398). The constant was chosen so that the nominal excluded volume for $CH_4$ is 1.0. So, molecular superpositions that occupy small volumes score high, and those occupying larger volumes score low. The overall objective function that drives the superposition algorithm is the mean pairwise similarity of a set of molecules in their alignment multiplied by this factor. A restraint against ligand self-clashing, taken from Surflex-Dock, is imposed during gradient-based conformational optimization.

**Procedure.** The following describes the overall procedure for generating an optimal superposition. There are two phases to the algorithm:

(1) Pairwise superposition: The best pairwise superpositions of all input ligands are generated. The resulting set of high-scoring superpositions is input to the next phase of the algorithm.

(1.1) Input: (a) M input ligands that are assumed to be protonated corresponding to aqueous solution at physiological pH. (b) Search depth to control the degree to which conformations of whole molecules are sampled and to control the degree to which conformations of molecular fragments are sampled (default 20).

(1.2) Output: A list of the $N$ best (default 100) pairwise superpositions, as scored by the objective function, which includes a term that minimizes volume.

(1.3) Procedure:

(1.3.1) All input ligands are searched to some maximum number of diversely chosen conformations (default 20). Each conformation of each ligand will be used as a target to which all other ligands are flexibly aligned, to maximize their molecular similarity.

(1.3.2) For each conformational target from 1.3.1, each molecule is flexibly aligned to that target to maximize molecular similarity. The score of the superposition is computed according to the objective function.

(1.3.3) Each result of 1.3.2 is added to a list of the $N$ best scoring (default 100) partial superpositions.

(2) Hypothesis augmentation: The partial hypotheses that serve as input to this step are augmented by addition of a single new molecule. The procedure maintains a constant list of the $N$ best hypotheses of a given length (default 100) and repeats until all hypotheses contain poses for all input ligands.

(2.1) Input: Set of $N$ partial hypotheses each containing $Q$ aligned molecules.

(2.2) Output: The $N$ best hypotheses each containing $Q + 1$ molecules.

(2.3) Procedure (for each hypothesis in the input set):

(2.3.1) For each aligned molecule of the current hypothesis, each molecule not already in the hypothesis is flexibly aligned to maximize molecular similarity. The score of the superposition is computed according to the objective function.

(2.3.2) Each of the resulting hypotheses (now $Q + 1$ aligned molecules) is added to a list of the N best hypotheses of length $Q + 1$.

(2.3.3) The hypothesis augmentation step is repeated until $Q + 1$ is equal to $M$ (i.e. all input ligands are present in all $N$ hypotheses). The process typically takes from a few to several hours of real time on standard desktop computers.

Given a molecular superposition, which will also be referred to later as a model or binding site hypothesis, the procedure for computing the score of a new ligand is as follows:

(1) Input: (a) $M$ ligands comprising the model. (b) A new ligand, protonated per expectation at physiological pH.

(2) Output: (a) a score between 0 and 1, and (b) the pose of the input ligand that gives rise to the reported score.

(3) Procedure:

(3.1) For each of the $M$ ligands in the model:

(3.1.1) The input ligand is flexibly aligned to the model ligand to maximize similarity to the model ligand, resulting in 10 poses.

(3.1.2) For each of the 10 optimized poses of the input ligand:

(3.1.2.1) The similarity score to each of the $M$ model ligands is computed.

(3.1.2.2) The mean score is computed.

(3.1.2.3) If the mean score exceeds the current maximum mean score, the maximum is modified and the ligand pose is saved.

(3.2) The maximum mean score is returned along with the corresponding pose.

So, the score of a new ligand is intended to reflect its ability to mimic, in a single pose, the model represented by the joint superposition of $M$ molecules. No penalty is imposed on excessive volume, since it would amount only to a bias against large molecules.

**Computational Procedures.** Following preparation of the molecular data, as described above, Surflex-Sim was run to generate molecular superpositions for the five test cases described ("surflex-sim -maxconfs 100

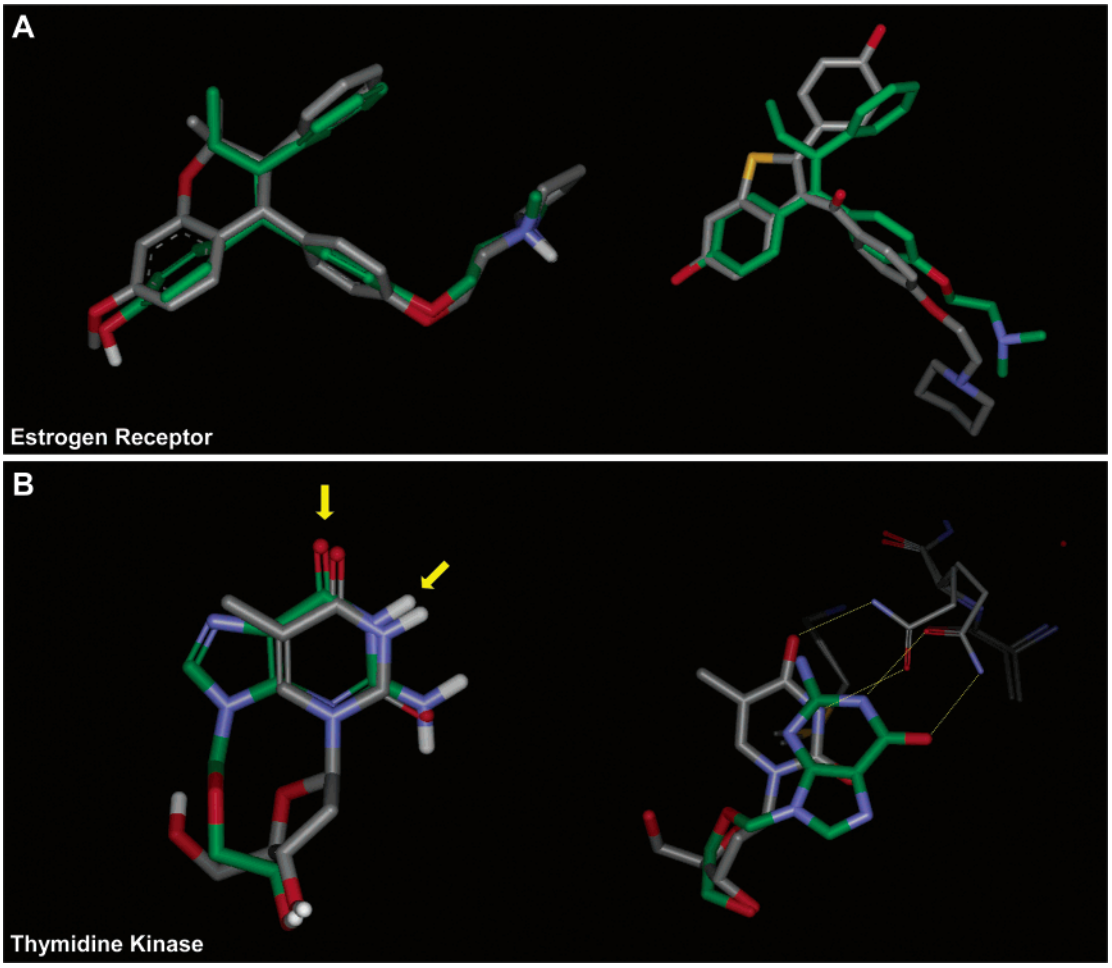


**Figure 5.** Panel A: Optimized molecular superpositions for ER ligands 4-hydroxytamoxifen and centchroman (left) and experimentally derived relative alignment of 4-hydroxytamoxifen and raloxifene (right). Panel B: Optimized molecular superpositions for TK ligands deoxythymidine and acyclovir (left) and crystallographic relative alignment of the same two ligands with protein residues Gln125, Ala124, and Met128 (right). In the ER case, judging by analogy to raloxifene, which is structurally related to centchroman, the correct correspondence of structural ligand features was achieved, with the unconstrained pendant ether substituent varying in absolute configuration from experiment. In the TK case, the induced relative alignment of the ligands is clearly incorrect, in a direct sense. However, since Gln125 makes a substantial change on binding pyrimidines versus purines, the correspondence of hydrogen bonding features in the induced alignment is correct. The indicated carbonyl oxygen and proton on the left are making the same contacts on the right. Note: nonpolar hydrogens have been suppressed in the similarity-based superpositions for clarity, but they are used in the underlying computation.

hypo InputMoleculeList log"). For each case, this resulted in up to 100 scored superpositions. The top five scoring were further optimized by gradient descent, and the highest scoring hypothesis was selected, subject to the constraint that the reported level of ligand internal clashing was not substantially higher than the other high-scoring hypotheses. The final molecular superpositions for each ligand set were then treated as quantitative models of the binding determinants of the respective proteins. To evaluate the utility of these models, known ligands and random ligands were scored for their degree of fit, using similarity optimization. The score of a ligand against a model was simply the maximum (over different poses of the ligands) mean similarity of a single pose of the ligand to the individual molecules comprising the model. So, the scores reflect the extent to which a ligand could best mimic the joint superposition of molecules within a model. The 2D graph-based Tanimoto molecular similarity method was used as a control for purposes of comparison (as implemented in Cambridge Software's ChemFinder program).[25]

## Results and Discussion

Clearly, the goal of methods to superimpose ligands is 2-fold: (1) to make an accurate prediction of the relative conformation and alignment of the ligands as they are bound in the native protein; and (2) to generate models that can be exploited for computational assessment of potential candidate molecules. In what follows, the ER and TK data sets described earlier will serve to address both issues and provide a direct comparison with docking methods. The GPCR and $GABA_A$ receptor ligand-based models will be used to probe the second issue both with respect to utility in screening databases of compounds and in terms of selectivity among related receptors. The serotonin case will also be used to consider whether direct prediction of relative potency is possible using such sparse models.

**Comparison to Docking-Based Methods: ER and TK.** Figure 5 shows the optimal superpositions of ligands for the ER and TK proteins (see the boxed compounds from Figures 2 and 3) along with relative crystallographic alignments. In the ER case, a structure

**Table 2.** Comparative False Positive Rates for Screening: Similarity versus Docking (note that true positive rate (TP) of 80% Means 8/10 for the Docking Approaches, but 7/9 and 6/8 for the ER and TK Similarity Cases, Respectively)

| | false positives from random ligands, % | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | thymidine kinase | | | | | estrogen receptor | | | | |
| TP, % | Surflex-Sim | Surflex-Dock | DOCK | FlexX | GOLD | Surflex-Sim | Surflex-Dock | DOCK | FlexX | GOLD |
| 80 | 0.3 | 0.9 | 23.4 | 8.8 | 8.3 | 0.0 | 1.3 | 13.3 | 57.8 | 5.3 |
| 100 | 0.7 | 3.2 | 27.0 | 19.4 | 9.3 | 13.4 | 2.9 | 18.9 | | 23.4 |

of centchroman bound to ER was not available, so raloxifene was used as a surrogate. The predicted correspondence on the left is mirrored quite nicely by experiment, with the notable exception that the pendant functionality of the ether, which is important in binding, does not have the correct absolute conformation. This is a limitation of ligand-based methods; where flexibility among several ligands occurs in the same place, there is no constraint on the induced overlap. Surflex-Sim in such cases will be biased based on minimizing the overall volume. Automatic detection of such cases with explicit modeling of the pendant substituents is an area for further exploration.

In the case of TK, the relative alignment of the two ligands is clearly incorrect with respect to the overlap of the heterocycles. However, there is a subtlety with respect to conformational adaptation on the part of TK in binding different classes of nucleoside analogues: Gln125 is flipped in orientation and shifted in binding acyclovir versus deoxythymidine. The indicated carbonyl oxygen and proton in the induced alignment each make contact with the *same* heavy atoms of the protein. So, in a sense, the induced alignment on the left is a reasonable approximation to the actual alignment on the right. However, while the correspondence of the indicated atoms is appropriate, the tight spatial correspondence of the other carbonyl and the $NH_2$ is wrong. This illustrates another limitation of ligand-based methods: given a case where a protein active site undergoes significant conformational adaptation on binding different ligands, particularly similar ones, it is generally not possible to induce a geometrically accurate superposition of ligands since the underlying assumption is that the ligands are binding to highly similar sites. This is not to say that methods dependent on alignments for making quantitative predictions will necessarily be inaccurate, just that certain predictions predicated on the juxtaposition of two molecules whose cognate binding pockets are different may be inaccurate.

Given that the geometric overlaps in the ER and TK cases were only correct to a degree, it is interesting to consider how well the two molecular superpositions, when used as binding site models, ranked the remaining known ligands among the random screening compounds (nine actives for ER, eight for TK, with 941 random compounds screened against each). For the ER case, the cognate ligands were ranked: 1−7, 62, and 135 (Figure 2). For the TK case, the cognate ligands were ranked: 1, 5−10, and 15 (Figure 3). Table 2 shows the false positive rates corresponding to 80% and 100% true positives along with the results for multiple docking techniques.[10,11] With the exception of Surflex-Dock's performance on ER at the 100% TP level, Surflex-Sim performs uniformly better than the docking techniques in both the TK and ER cases.

It is very important to reiterate that this is *not* a fair direct comparison to the docking methods, due to the problem of inductive bias. Whereas the docking methods are essentially parameter free and have no knowledge of the ligands under study, the construction of ligand-based models embeds knowledge of at least some structural types of true ligands. In the TK case, where the ligands used to construct the model contain the same two heterocycles as all of the test ligands, the Surflex-Sim approach does extremely well. However, it is likely that even very naïve methods would perform well, given information about the two major structural types of ligands. Still, it is interesting that the nominally inaccurate geometric overlap was not an important issue in achieving low false positive rates. In the ER case, of the nine ligands used for testing only three (nafoxidine, LY-326315, and raloxifene) might be considered highly structurally similar to either of the two parent compounds. Of the remaining six, only the bottom two from Figure 2 were ranked out of the top group. The remaining four ligands (EM-343, Sumitomo biphenol, ZK-11901, and LY-357489) represent different chemotypes from the two molecules used to construct the model and were correctly recognized as being highly similar to the induced model.

**GPCR and $GABA_A$ Models.** A more relevant situation for drug discovery is represented by targets such as GPCRs and ion channels, where structure determination is challenging. To explore these protein classes, four models were constructed, representing putative binding site ligand geometries for the serotonin, muscarinic, histamine, and $GABA_A$ receptors (see Figure 1 for the ligands used for model construction). For each ligand set an optimal superposition was generated by Surflex-Sim using identical parameters and procedures (as detailed above).

Figure 6 shows the final models for the serotonin, muscarinic, and histamine receptors. The independently derived models are oriented to reflect their remarkable degree of coincidence. The right-hand sides are the bottom view of the respective left-hand sides. In the case of the molecules A and B from Figure 1 (the 5-$HT1_A$ ligands), the resulting alignment perfectly superimposed the positively charged amine protons in terms of both position and orientation. The oxygens are both capable of accepting hydrogen-bonds from protons in a single location, and the hydrophobic envelopes of the molecules are remarkably similar. This is almost identical to the alignment that resulted in a highly predictive quantitative model of multiple 5-$HT1_A$ chemotypes using the Compass method.[9] However, the Compass work utilized somewhat ad hoc methods for generating superpositions, and it took *much* longer than the current approach.

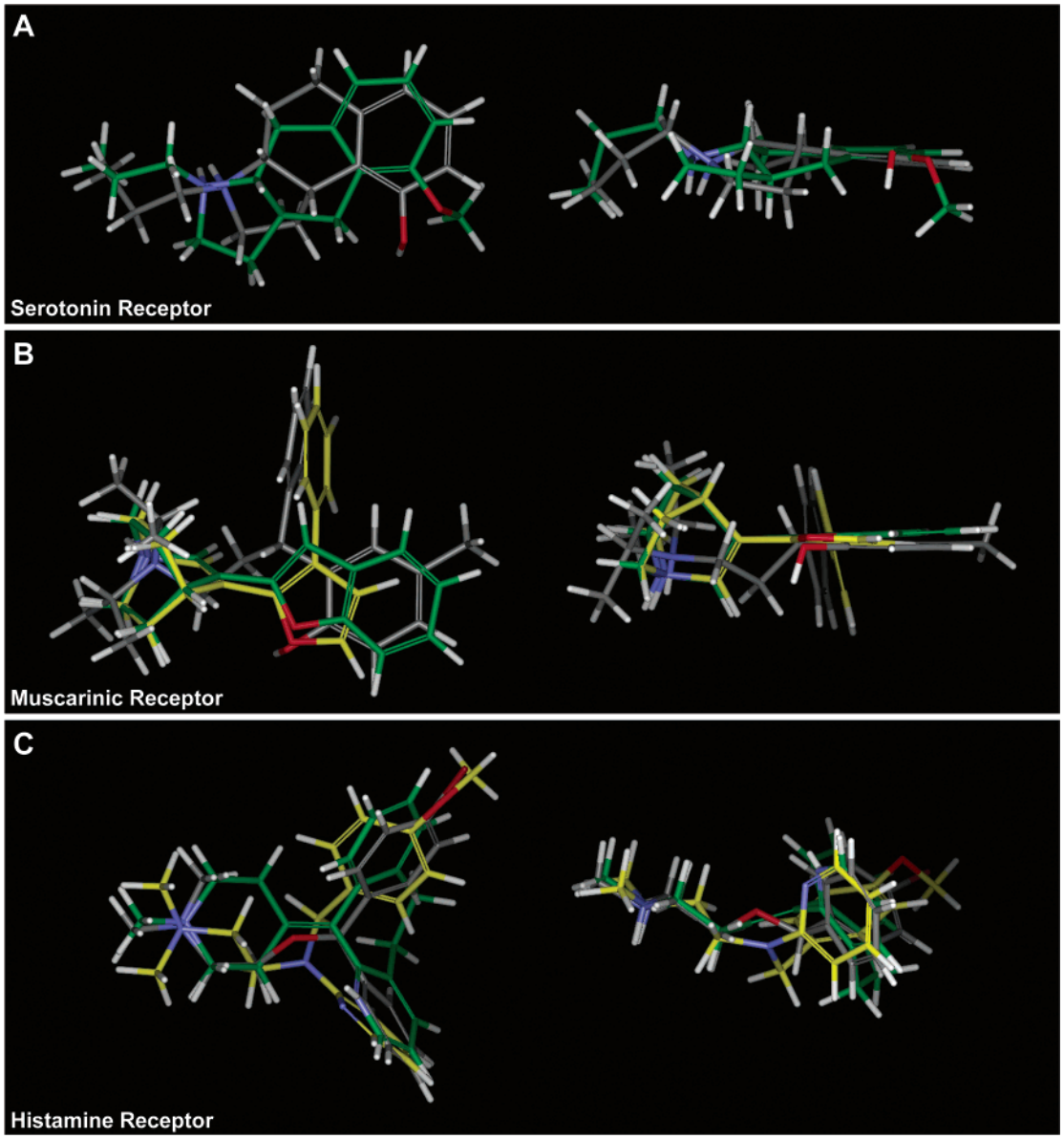


**Figure 6.** Optimized molecular superpositions for serotonin, muscarinic, and histamine receptor ligands (panels A, B, and C, respectively). Molecule carbons are colored according to column position from Figure 1 (first = gray, middle = yellow, last = green). Views on the right are rotated to show the left views from the bottom.

In the case of the muscarinic antagonists (molecules C−E from Figure 1), significant molecular flexibility of tolterodine coupled with very different underlying chemical scaffolds combine to present a more challenging case. Panel B of Figure 6 shows the highest-scoring hypothesis, which is remarkable in that such different chemotypes can be so convincingly superimposed. As with the 5-HT1$_A$ case, the polar and hydrophobic moieties of the molecules agree in position, orientation, and volume occupancy. The substituted furan (molecule D) and benzofuran (molecule E) each account for different hydrophobic aspects of tolterodine (C).

In the case of the first-generation histamine antagonists (molecules F−H from Figure 1), we see a superposition that is reminiscent of the model for the antimuscarinics. This is not terribly surprising, since many compounds are known to antagonize both receptor types. Of the 44 known antimuscarinics and 48 antihistamines used to test the models, 28 are annotated as having effects against both types of receptors. All three GPCR ligand-based models exhibit similar structures, with a requirement for a protonated amine, preference for a proton acceptor, and variable requirements for hydrophobic shape. Note, however, that the detailed orientation of the generally hydrophobic portions of the ligands have somewhat different relative orientations among the different models.

The GABA$_A$ receptor benzodiazepine binding site model (constructed from molecules I−K) has different requirements, which are illustrated in Figure 7. The overlay of diazepam and alprazolam (molecules I and J) is uncontroversial. However, it is not clear that the superposition of zopiclone (molecule K) is obviously correct, though it does largely agree with previously published models of the benzodiazepine binding site pharmacophore.[26] Exploration of GABA$_A$ receptor interactions with classic benzodiazepine agonists versus other chemotypes using photoincorporated ligands suggests that conformational adaptation may play an important role in recognition of ligands in this recep-

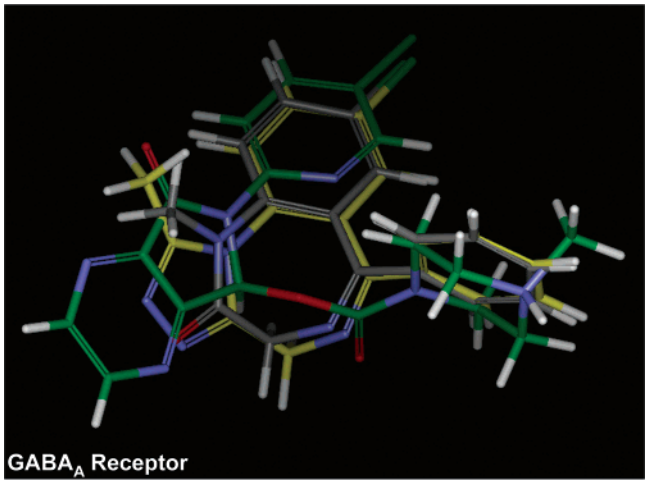


**Figure 7.** Optimized molecular superposition for the benzodiazepine binding site of the GABA$_A$ receptor (molecules I−K from Figure 1). Carbons are colored as in Figure 6.

tor.[26,27] So, the observations above with respect to thymidine kinase may also be relevant in this case.

**Quantitative Evaluation of Structural Hypotheses.** While Figures 6 and 7 offer some support that the computational procedures produced intuitively sensible results, quantitative evaluation of the models is critical. Figure 8 shows ROC curves comparing the performance of Surflex-Sim to the 2D-based Tanimoto similarity method for each of the ligand-based hypotheses, illustrating enrichment of cognate ligands over random screening molecules. All of the known and random ligands were processed, aligned, and scored to the models using identical procedures and parameters for Surflex-Sim. Procedures for the 2D method were also uniform, and the method performed best by using maximum score over the model ligands. For the serotonin, muscarinic, and histamine receptor cases, Surflex-Sim performed uniformly better than the 2D method. Given a fixed true positive rate for one method, we can compare the methods by considering the true positive rate for the other method at an equivalent *false positive* rate. The true positive rates for the 2D method corresponding to 90% TP for Surflex-Sim were 72%, 20%, and 22%, respectively, for the serotonin, muscarinic, and histamine receptor cases.

In the GABA$_A$ case, where 9/15 of the test ligands had classic benzodiazepine scaffolds (as did 2/3 of the molecules for model induction), the 2D method was able to achieve 100% TP with Surflex-Sim at 90%. Overall performance for the GABA$_A$ case was quite similar for the two methods. Surflex-Sim was able to retrieve all nine test benzodiazepines with no false positives compared with 15/850 false positives for the 2D method (1.8%). The binding of GABA$_A$ ligands to the benzodiazepine binding site has been extensively modeled by other researchers; Huang et al.[28] have made a comprehensive study and suggest multiple binding modes for different classes of ligands. While the superposition produced by Surflex-Sim was similar to that illustrated by Davies et al.,[26] the six test molecules that were not benzodiazepines may bind differently. The Surflex-Sim model makes a strong implicit assumption of similarity of binding mode, which complicates retrieval in cases where the assumption may be false. There was a sharp drop-off in scores going from the lowest scoring benzodiazepine (quazepam 0.78) to the highest scoring non-benzodiazepine (zolpidem 0.67).

For Surflex-Sim, 60% recovery of known cognate ligands resulted between 0 and 3% of the random ligands being found as false hits. For 70% and 80% recovery, the respective false positive rates were 2−6% and 3−11%. The recently published fFLASH method[29] reported nominally comparable results on a screening benchmark (3% FP rate for a 50% TP rate) using the structure of folate bound to DHFR as an alignment target. However, 50/51 of the test molecules were quinazoline-based DHFR inhibitors, which presented very little significant structural divergence from the parent molecule.

The problem of inductive bias is less of an issue with the ligand-based models discussed here, compared with either the TK/ER cases or the fFLASH study. The molecules chosen for model construction were either from the limited series of single medicinal chemistry groups (serotonin and muscarinic) or are very old examples of ligands in their target classes (histamine and GABA$_A$). Consequently, a large number of the test ligands represent different structural types than those used for model construction (though this was less true for the GABA$_A$ receptor ligands). Figure 9 shows the structures of selected cognate ligands that scored very high relative to random molecules for each of the four models. In the serotonin and muscarinic cases, there are a number of novel ligands that are recognized as very high scoring which have little obvious structural similarity to the ligands used for model induction. Using the respective background score distributions, it is possible to compute a percentile-based score for the ligands, which are shown under the ligand names. Six of the twelve molecules shown are within the top 1% of random scores for each respective model. Ten of the twelve are within the top 2%, with only zolpidem (96.8) and zaleplon (87.1) being lower. Clearly, molecules with diverse and novel scaffolds can be distinguished from random molecules based on small numbers of known ligands.

While the notion of structural similarity of scaffolds is somewhat subjective, it is probably safe to characterize the ligands represented in Figure 9 as being quite substantially more different from the corresponding molecules in Figure 1 than was the case for the ER and TK model induction cases. Given that the issue of inductive bias is somewhat reduced compared with the results presented above, we see that across all four pure ligand-based cases, a false positive rate of 0−3% corresponded to a true positive rate of about 60%. Recall the results from Table 2 for ER and TK (which depended on the same population of random ligands) that the best of the docking methods achieved roughly 80% true positive rates at a similar level of false positives. The next best docking method yielded about 5−8% false positives for an 80% true positive rate, which is quite similar to what was observed for Surflex-Sim in these four cases (3−11%). So, while it is clearly possible to achieve better performance with structure-based techniques, depending on the docking method chosen, a pure ligand-based approach such as Surflex-Sim can achieve competitive screening performance.

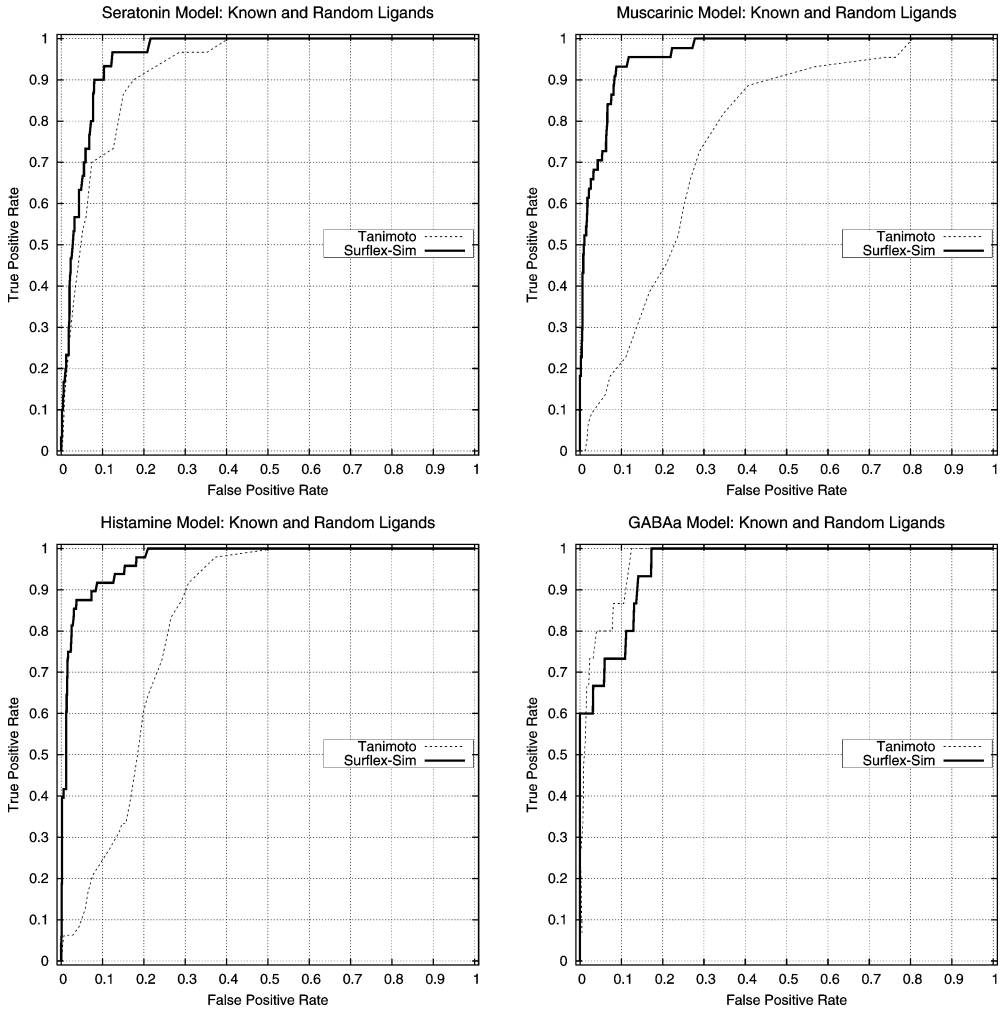


**Figure 8.** ROC curves GPCR and $GABA_A$ receptor models (solid lines for Surflex-Sim and dotted lines for the 2D Tanimoto similarity method). In each case, for both methods, there is a highly significant enrichment of cognate over random ligands ($p \ll 0.001$ in all cases by ROC area). For the serotonin, muscarinic, and histamine cases, Surflex-Sim performs significantly better than the 2D method. In the $GABA_A$ case, where a large proportion of the test ligands were of highly similar structural similarity to the molecules used for model construction, there is little difference in performance.

Theoretical enrichment rates (the fold excess of observed hits to expected hits given a selected subset of a library) are computable from the data presented above. They are dependent on the proportion of the library chosen for screening, which is based on the score threshold applied to define the subset. With a library of $(N + M)$ molecules where $M$ are true hits and $N$ are not ($N \gg M$), we choose a score threshold T that picks some proportion of molecules of the $(N + M)$. If we make the assumption that the distribution of scores for true positives and random compounds is close to the distributions that underlie the ROC computations, we will observe the computed true positive rate (TP) and false positive rate (FP) at threshold T. So, we will observe $TP \cdot M$ hits from a total number of tested compounds equal to $TP \cdot M + FP \cdot N$. Our expected number of hits is then $((TP \cdot M + FP \cdot N)/(N + M)) \cdot M)$. Our enrichment rate is the ratio of the actual number of hits to the expected number: $TP/((TP \cdot M + FP \cdot N)/(N + M))$. For $N \gg M$ (e.g. 100 000 nonhits and 50 hits), this simplifies to TP/FP. With score thresholds that select a large proportion of libraries, the enrichment rate will tend toward 1.0, since both the TP and FP rates will tend toward unity (upper right corner of ROC plots). But with very small proportions of a library selected based on a high score threshold (left edge of ROC plots), high enrichment is possible. For the muscarinic and histamine models, at score thresholds that select about 0.2−0.5% of a library, we observed theoretical enrichment rates slightly over 150-fold in both cases. For the $GABA_A$ case, which was skewed by the presence of a large number of analogues of the modeled ligands, the enrichment rate was higher (over 250-fold). For the serotonin case, which was built on just two ligands of vastly different structure than the test ligands, the enrichment rate was lower (about 50-fold). These observations are in rough agreement with a published application of Surflex-Sim in screening for inhibitors of PARP (poly ADP ribose polymerase).[30] In that application, a single rigid compound was used as a target for similarity-based screening (obviating the need for construction of a multi-ligand hypothesis), and 6 of 22 selected compounds from a large screening library were shown to be specific inhibitors of PARP. The library's proportion of true positives was 10/16 000 (determined by high-throughput screening), and the corresponding enrichment rate was therefore about 450-fold.

**Selectivity and Potency.** The foregoing has addressed a more or less binary characterization of potency and selectivity: the difference between known ligands

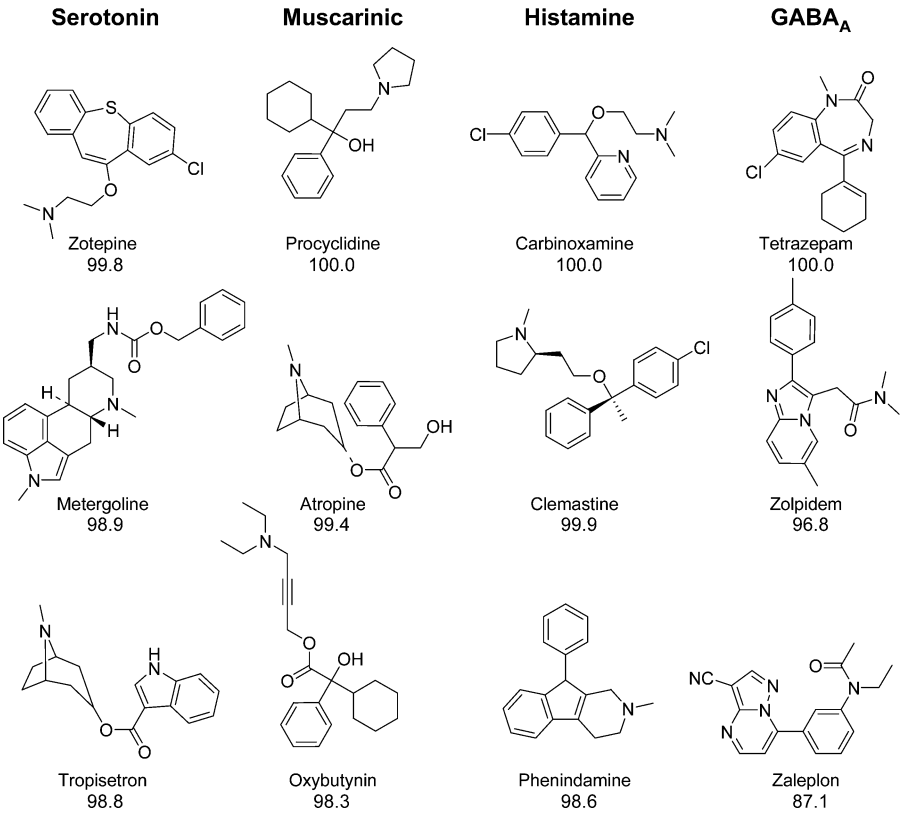


**Figure 9.** Examples of high-scoring ligands against each of the four ligand-based models. Under the name of each ligand is the percentile rank of its score relative to random molecules within its cognate model.

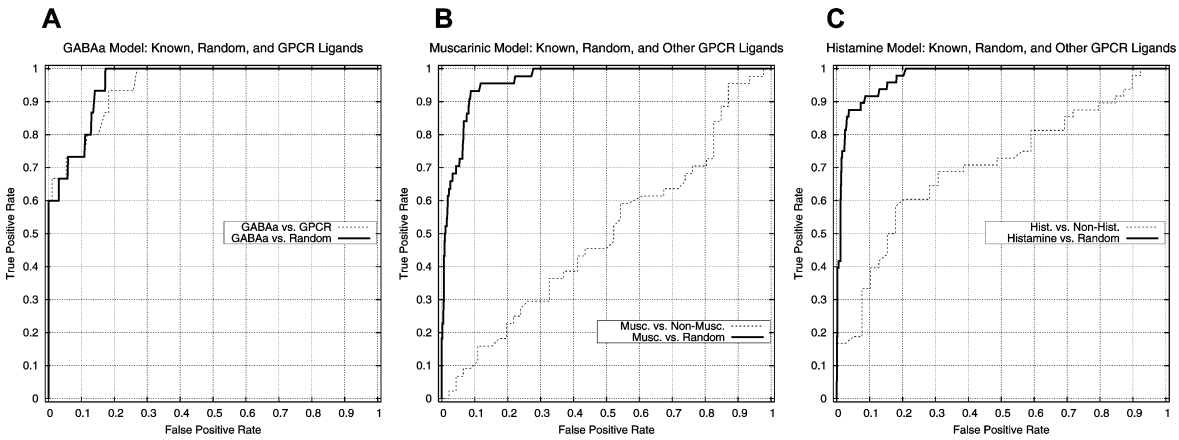


**Figure 10.** ROC curves illustrating selectivity for GABA$_A$, histamine, and muscarinic models (solid lines for cognate versus random compounds and dotted lines for cognate versus other decoy sets).

of a protein and ligands thought not to bind at all. In that domain, the ligand-based structural hypotheses are effective, both qualitatively with respect to plausibility of molecular superpositions as well as quantitatively in virtual screening. However, it is useful to consider to what degree such simple models can address more subtle distinctions involved in ligand binding. Figure 10 shows ROC plots muscarinic, histamine, and GABA$_A$ models that include enrichment versus noncognate ligands from the GPCR ligand set of the respective receptor model. Plot A shows that GPCR ligands and random screening ligands both had very similar separation from the cognate ligands against the GABA$_A$ model. This observation should ease concerns that the random screening library may be biased toward "nondruglike" molecules since it cannot be argued that the GPCR ligands are not druglike, as most of the examples are human or animal therapeutics. Note also that the GABA$_A$ model included a ligand with a protonated tertiary amine, which would have enriched for high scores among ligands of the other three receptors.

Differences *among* GPCR ligands are much more subtle than differences *between* GPCR ligands and other drugs. Plot B of Figure 10 shows no separation between the scores for muscarinic ligands and the remaining GPCR ligands not annotated as having muscarinic effects. Both annotated muscarinic receptor ligands and those with no annotated muscarinic effects fit the model equally well, with the exception that the subset of other GPCR ligands with annotated serotonin receptor effects was slightly separated (data not shown). By contrast, the histamine model, shown in Plot C, makes a signifi-

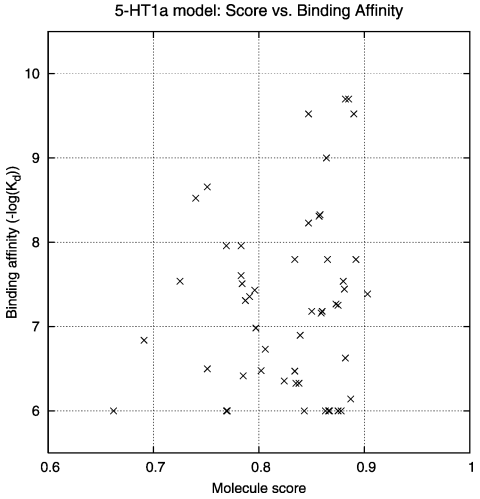


**Figure 11.** Plot of model score versus 5-HT$1_A$ binding affinity for 53 novel molecules from six molecular series.

cant separation between ligands of the histamine receptor and those GPCR ligands not thought to bind the histamine receptor. It is quite possible that the muscarinic model is inadequate, either based on the objective function used to choose the optimal model, the method of scoring new molecules against the model, or based on the choice of molecules from which to induce the model (which represent smallish muscarinic antagonists). However, since it is likely that the annotation data that underlies the molecular classes is incomplete, some proportion of molecules not listed as having muscarinic effects, in fact, may have such effects. As mentioned earlier, a large proportion of muscarinic ligands in the testing set are also identified as having histamine effects, and vice versa. Since one of the primary side-effects of anithistamines derive from muscarinic origins (dry mouth, blurred vision, and dysfunctional urine voiding),[31] it appears that ligands with primarily antihistaminic effects are a priori likely to affect muscarinic receptors. The primary data source used to derive annotations was GPCRDB,[22] and there are a number of examples of histamine receptor ligands lacking specific binding data for muscarinic receptors, but which are known to have such effects based on pharmacology. For example, all three molecules used for histamine model construction (bromodiphenhydramine, pyrilamine, and azatadine) have side-effects related to muscarinic activity. Several molecules used for model testing (e.g. hydroxyzine, astemizole, carbinoxamine, and triprolidine) also have documented muscarinic pharmacological effects, despite lacking such annotation in GPCRDB. However, similar inspection of the muscarinic ligands does not reveal histaminergic pharmacological effects. So the apparent lack of specificity in the muscarinic model *may* reflect biological reality. Regardless, the selectivity exhibited by the histamine model in distinguishing among different classes of GPCR ligands is encouraging.

Quantitative prediction of potency using 3D approaches begins from an alignment such as the ones produced here. In the case of the serotonin model, there are an additional 53 molecules in addition to the two used for construction of the serotonin model whose binding affinity to the 5-HT$1_A$ receptor is known.[9] Figure 11 shows the relationship between the score using the simple two-molecule model and the measured p$K_d$ of these 53 molecules. The data exhibit a lower right triangular shape. Molecules that are very active tend to have high scores (5/5 with p$K_d \geq 9.0$ have scores $\geq 0.85$, $p = 0.03$ by exact binomial); molecules that are inactive have lower scores than the active group (scores of molecules with p$K_d \leq 7.0$ vs p$K_d \geq 9.0$, $p \ll 0.01$ by *t*-test); molecules that have high scores have widely varying binding affinities (p$K_d$ of molecules with scores $\geq 0.85$ vs $< 0.85$ are not different by *t*-test). This is not surprising, since the model was constructed only from active molecules. So, while the model is able to capture some aspects of what is required for binding, it is unable to induce constraints regarding detriments to binding, and thus many high-scoring ligands have poor binding affinity. Previous work with Compass[7] on this data set validates the two-molecule hypothesis, since it is nearly identical to the overlay that resulted in a quantitatively predictive model of binding affinity. Compass and related methods build upon the initial alignments in large part by learning the excluded volume of interaction, essentially finding the "walls" of a protein binding site.

## Conclusions

The work presented here offers a generally applicable method for producing ligand-based binding site hypotheses, which can be used directly for high-throughput virtual screening or to form the basis on which to construct more detailed models of molecular activity. The algorithms implemented within Surflex-Sim are capable of addressing molecules with conformational flexibility that is typical of what is found in therapeutically interesting small molecules. Performance in terms of screening utility is comparable to that of many structure-based molecular docking techniques, but the best docking methods are capable of better sensitivity and specificity. From small sets of known ligands, it is possible to induce molecular superpositions that form the basis for predictive models of molecular activity. These models are sufficiently accurate and scaffold-independent that structures with widely varying chemotype can be recovered from screening libraries at very low false positive rates. Theoretical enrichment rates of 150-fold over random screening are possible using just three ligands for model construction.

There are a number of purely methodological improvements that will be a focus of future work, including more efficient hypothesis generation, development of means by which to objectively select from multiple hypotheses which score equally well, and full induction of quantitatively predictive models of activity. However, given that the methods presented are general and fully automatic, direct application of the existing methods in high-throughput to construct models of large numbers of therapeutically relevant targets will be a high priority. It should be possible to enable rapid virtual screening against many tens of biological targets, which might prove to be of use in suggesting potential side-effect modulators of molecules undergoing development toward clinical application. Such an effort will also produce a large set of benchmarks for quantitative evaluation of many problems in the 3D QSAR field.

**Acknowledgment.** The author is grateful to Ann E. Cleves for help with data curation and for comments on the manuscript and to NIH/NCI for partial funding of the work (CA64602).

## References

(1) Walters, P. W.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(2) Renfrey, S.; Featherstone, J. Structural proteomics. *Nat. Rev. Drug Discovery* **2002**, *1*, 175−176.
(3) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroid to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(4) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.
(5) Willett, P. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J. Mol. Recognit.* **1995**, *8*, 290−303.
(6) Clark, D. E.; Westhead, D. R.; Sykes, R. A.; Murray, C. W. Active-site-directed 3D database searching: pharmacophore extraction and validation of hits. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 397−416.
(7) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. A shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.
(8) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(9) Jain, A. N.; Harris, N. L.; Park, J. Y. Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT1A receptor. *J. Med. Chem.* **1995**, *38*, 1295−1308.
(10) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(11) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(12) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46*, 2895−2907.
(13) Hindle, S. A.; Rarey, M.; Buning, C.; Lengauer, T. Flexible docking under pharmacophore type constraints. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129−149.
(14) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.
(15) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein−ligand recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199−213.
(16) Lin, C. H.; Haadsma-Svensson, S. R.; Lahti, R. A.; McCall, R. B.; Piercey, M. F.; Schreur, P. J.; Von Voigtlander, P. F.; Smith, M. W.; Chidester, C. G. Centrally acting serotonergic and dopaminergic agents. 1. Synthesis and structure−activity relationships of 2, 3, 3a, 4, 5, 9b-hexahydro-1H-benz[*e*]indole derivatives. *J. Med. Chem.* **1993**, *36*, 1053−1068.
(17) Lin, C. H.; Haadsma-Svensson, S. R.; Phillips, G.; Lahti, R. A.; McCall, R. B.; Piercey, M. F.; Schreur, P. J.; Von Voigtlander, P. F.; Smith, M. W.; Chidester, C. G. Centrally acting serotonergic and dopaminergic agents. 2. Synthesis and structure−activity relationships of 2, 3, 3a, 4, 9, 9a-hexahydro-1H-benz[*f*]-indole derivatives. *J. Med. Chem.* **1993**, *36*, 1069−1083.
(18) Nilvebrant, L. Tolterodine and its active 5-hydroxymethyl metabolite: pure muscarinic receptor antagonists. *Pharmacol. Toxicol.* **2002**, *90*, 260−267.
(19) Nordvall, G.; Sundquist, S.; Johansson, G.; Glas, G.; Nilvebrant, L.; Hacksell, U. 3-(2-Benzofuranyl)quinuclidin-2-ene derivatives: novel muscarinic antagonists. *J. Med. Chem.* **1996**, *39*, 3269−3277.
(20) Johansson, G.; Sundquist, S.; Nordvall, G.; Nilsson, B. M.; Brisander, M.; Nilvebrant, L.; Hacksell, U. Antimuscarinic 3-(2-furanyl)quinuclidin-2-ene derivatives: synthesis and structure−activity relationships. *J. Med. Chem.* **1997**, *40*, 3804−3819.
(21) O'Neil, M. J.; Smith, A.; Patricia, E. H.; Obenchain, J. R.; Gallipeau, J. A. R.; D'Arecca, M. A. *Merck Index: An Encyclopedia of Chemicals, Drugs, & Biologicals*, 13th ed.; Merck & Co.: Rahway, NJ, 2001.
(22) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **1998**, *26*, 275−279.
(23) Feher, M.; Schmidt, J. M. Multiple flexible alignment with SEAL: a study of molecules acting on the colchicine binding site. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 495−502.
(24) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505−1514.
(25) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(26) Davies, M.; Newell, J. G.; Derry, J. M.; Martin, I. L.; Dunn, S. M. Characterization of the interaction of zopiclone with $\gamma$-aminobutyric acid type A receptors. *Mol. Pharmacol* **2000**, *58*, 756−762.
(27) McKernan, R. M.; Farrar, S.; Collins, I.; Emms, F.; Asuni, A.; Quirk, K.; Broughton, H. Photoaffinity labeling of the benzodiazepine binding site of $\alpha1\beta3\gamma2$ $\gamma$-aminobutyric acid A receptors with flunitrazepam identifies a subset of ligands that interact directly with His102 of the alpha subunit and predicts orientation of these within the benzodiazepine pharmacophore. *Mol. Pharmacol* **1998**, *54*, 33−43.
(28) Huang, Q.; He, X.; Ma, C.; Liu, R.; Yu, S.; Dayer, C. A.; Wenger, G. R.; McKernan, R.; Cook, J. M. Pharmacophore/receptor models for GABA(A)/BzR subtypes ($\alpha1\beta3\gamma2$, $\alpha5\beta3\gamma2$, and $\alpha6\beta3\gamma2$) via a comprehensive ligand-mapping approach. *J. Med. Chem.* **2000**, *43*, 71−95.
(29) Kramer, A.; Horn, H. W.; Rice, J. E. Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 13−38.
(30) Perkins, E.; Sun, D.; Nguyen, A.; Tulac, S.; Francesco, M.; Tavana, H.; Nguyen, H.; Tugendreich, S.; Barthmaier, P.; Couto, J.; Yeh, E.; Thode, S.; Jarnagin, K.; Jain, A. N.; Morgans, D.; Melese, T. Novel inhibitors of poly(ADP-ribose) polymerase/PARP1 and PARP2 identified using a cell-based screen in yeast. *Cancer Res.* **2001**, *61*, 4175−4183.
(31) Hardman, J. G.; Limbird, L. E.; Gilman, A. G. *Goodman and Gilman's: The Pharmacological Basis of Therapeutics*, 10th ed.; The McGraw-Hill Companies, Inc.: New York, 2001.

JM030520F