JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

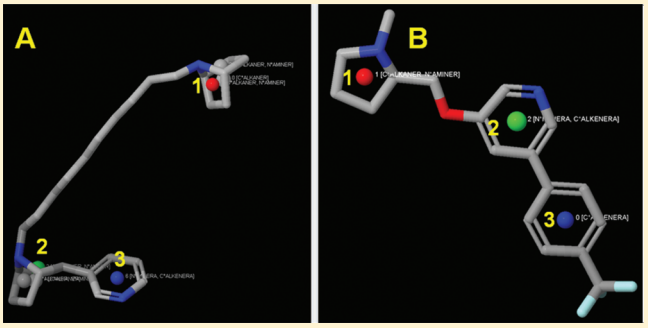
Fighting High Molecular Weight in Bioactive Molecules with Sub-Pharmacophore-Based Virtual Screening

Modest von Korff,*[,†] Joel Freyss,[†] Thomas Sander,[†] Christoph Boss,[‡] and Claire-Lise Ciana[‡]

[†]Department of Research Informatics and [‡]Drug Discovery Chemistry Actelion Pharmaceuticals Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

Supporting Information

ABSTRACT: A new subpharmacophore-based virtual screening method is introduced. Subpharmacophores are derived from large active molecules to detect small bioactive molecules as seeds for starting points in medicinal chemistry programs. A large data set was assembled from the ChEMBL database to check the validity of this approach. Molecules for 133 targets with molecular weights between 450 and 850 were selected as queries. For the query molecules, the pharmacophore descriptors were calculated. Up to 56 000 subpharmacophore descriptors with five to seven pharmacophore points were derived from the query pharmacophores. The subpharmacophore descriptors were used as queries to screen 1079 test data sets, containing decoys and spike molecules. A maximum upper molecular weight limit of 400 Da was set for the test molecules. Three different chemical fingerprint descriptors were used for comparison purposes. The subpharmacophore approach detected active molecules for 85 out of 133 targets and outperformed the chemical fingerprints. This ligand-based virtual screening experiment was triggered by the needs of medicinal chemistry. Applying the subpharmacophore method in a medicinal chemistry program, where a lead molecule with a molecular weight of 800 Da was available, resulted in a new series of molecules with molecular weights below 400.



1. INTRODUCTION

In mature lead-optimization programs, synthesized molecules tend to increase bioactivity together with molecular mass. Observed differences between compounds in early drug discovery programs and mature compounds were manifested in the rule of three[1] versus the rule of five.[2] An increase in mass worsens the desired physicochemical properties of the compounds (i.e., decreases the rate of diffusion through biological membranes exponentially with increasing mass).[3] This worsens the physicochemical properties needed for an orally available drug, such as the rate of absorption into the bloodstream and distribution in the target organ. In the past, the low molecular mass of a lead compound was often sacrificed to improve its bioactivity and selectivity.[4] To be able to assess the quality of a compound, under consideration of mass and bioactivity, ligand efficiency was introduced.[5,6] Ligand efficiency as a figure of merit is the quiet revolution in drug discovery by normalizing the potency of a compound having a number of "heavy" (non-hydrogen) atoms. Molecules leaving early drug discovery to enter the preclinical department are desired to show sufficient bioactivity and appropriate biophysicochemical properties.[7] Detecting such molecules in chemical space is the resulting challenge for medicinal chemists. Ligand efficiency, as an easy-to-calculate indicator, supports this optimization process. However multiobjective optimization requires a change in the paradigm that one scaffold is optimized until bioactivity is high enough. Inadequate physicochemical properties can cause a molecule series to drop out of a project. Hence, it is obvious that a new paradigm would require many more seed scaffolds for starting an optimization program than the old one did. These seed scaffolds would have to be delivered either from high-throughput screening or from the screening of focused libraries, the latter being more cost efficient. One possibility to generate focused libraries is ligand-based virtual screening. Known active molecules extracted either from the running medicinal chemistry program or from the literature are used as queries to search the compound libraries of the commercial suppliers or, if available, an in-house library. Query compounds originating from mature medicinal chemistry programs tend to be large and complex. However, using large and complex molecules to search commercial supplier databases with ligand-based virtual screening seldomly results in a satisfactory number of molecules, similar to the query. With an increased number of atoms, the chemical space is less and less densely populated. Taking this together with the undesired physicochemical properties of large molecules, there is a need to derive small, new bioactive molecules with diverse scaffolds from large query

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: August 29, 2011
Published: January 18, 2012

molecules.[8] This led to the idea of subpharmacophore screening. The idea that a pharmacophore is responsible for bioactivity is more than 30 years old: "A pharmacophore is an arrangement of molecular features or fragments forming a necessary but insufficient condition for biological activity."[9] Pharmacophore-based virtual screening started by encoding the knowledge that the reason for bioactivity of a chemical structure can often be reduced to some substructure patterns, with a certain 3D orientation toward each other.[10,11] Since then, several tools have been developed using the pharmacophore approach.[12−15] These tools take substructural features to define a three-dimensional pharmacophore model which may span the whole volume of the molecule. Using substructure-based pharmacophore models for virtual screening has raised some questions. If there is a large active molecule, can pharmacophore-based virtual screening be used to find small active molecules without *a priori* knowledge about which parts of the molecule are relevant pharmacophore points? Is there a pharmacophore described using a group of pharmacophore points, in which these are close enough together to be represented in a small molecule? What are the minimum and maximum numbers of pharmacophore points needed to detect bioactive molecules? These questions can be summarized into one: what is the minimum pharmacophore? Often it is not possible to derive pharmacophore models in a drug discovery project where crystal structures of the target are missing. However, without *a priori* knowledge about relevant pharmacophore points, a subset of small pharmacophore models can be generated. Instead of using one query pharmacophore for virtual screening, the database is searched using multiple subpharmacophore models. Small pharmacophore models demand different criteria for similarity calculation than larger ones. The probability of finding matching pharmacophores in the supplier database increases exponentially with a decreasing number of pharmacophore points, especially if these pharmacophore points differ only by a few Ångstroms. While this is, of course, an advantage for finding candidate molecules for biological screening, what is the probability of such molecules being active in the detection range of the assay? This question was already answered in the publication on ligand efficiency.[6] The calculation of ligand efficiency to assess the potential of a bioactive molecule in a medicinal chemistry program was driven by the experience that small molecules with low or moderate bioactivity have more potential to become the seed of a successful lead series than large molecules with high bioactivity. Conversely, decreasing the number of pharmacophore points just by one may result in a 10-fold loss of bioactivity. In addition, a matching set of pharmacophore points is not a guarantee that a molecule will be active. A molecule, chosen through virtual screening, will only be active if its pharmacophore can take the correct position to interact with the corresponding pharmacophore of the protein. And the conformational energy of the candidate molecule must be low enough not to overcompensate for the binding energy. As a result, a lower hit rate has to be expected, compared to other methods using complete query molecules. As is often the case in industrial pharmaceutical research, the subpharmacophore approach was initiated in response to a request out of a medicinal chemistry project. In a potent assay with limited throughput, a highly bioactive compound was found. Due to its molecular mass of 720 Da, medicinal chemists requested a compound with less molecular weight and good physicochemical properties. Thus, the next logical step was to use fragments of the lead molecule for virtual screening. The active compound was prepared through fragmentation in accordance with the RECAP rules.[16] Three chemical fingerprint descriptors were calculated for the fragments: FragFp, PathFp, and SphereFp. These fingerprint descriptors were proven to be well suited for enriching bioactive molecules in virtual screening experiments.[17] A library of commercially available compounds was virtually screened with the fingerprint descriptors derived from fragments of the bioactive compound. Since a high loss of activity was expected, the similarity criterion was set tight to an inverse Tanimoto of 0.9. Interestingly, no similar compound was found among the approximately 7 million supplier compounds. In a second virtual screening approach, sub-pharmacophore descriptors were used by applying the Flexophore descriptor implementation.[18] The pharmacophore points for the descriptor were calculated from the query fragments. The pharmacophore points were mapped on the Flexophore descriptor of the lead molecule and the corresponding distance histograms taken as edges for the sub-Flexophore descriptor. As a result, a sub-Flexophore descriptor was obtained with the same geometrical restrictions as the Flexophore descriptor derived from the whole molecule. Using the sub-Flexophore descriptors to search the supplier libraries resulted in approximately 400 hits. A total of 309 compounds were ordered from one supplier after some cherry picking by the medicinal chemists. In the biological testing that followed, a new chemical series was found and confirmed. All molecules in this series had molecular weights below 400. The best compound from this series showed the highest ligand efficiency in this project. Missing any reasonable virtual hit with the chemical fingerprints of the RECAP fragments triggered the consideration to use the sub-Flexophores without the detour via the RECAP fragments. Virtual screening with subpharmacophores while keeping the geometrical restrictions of the original molecule is the topic of this examination. To compile a test data set for subpharmacophore screening, the minimum requirements were defined as follows: For each target, at least one large ligand and 10 smaller ligands should be available. The large ligand would be used to generate the query descriptors, and the smaller ligands would be the spikes in the test data set. As the data source of choice, the ChEMBL database was selected.[19,20] It contains the bioactivity information of more than 500 000 molecules on more than 7000 targets. The database is freely available and can be downloaded in different database formats. As a source of decoys, a subset of the ChEMBL database was used. Each test data set contained 10 bioactive spike molecules and 990 decoys, selected randomly. Virtual screening using the test data sets was examined with the sub-Flexophores and three chemical fingerprint descriptors. Chemical fingerprint descriptors were used as a comparison method in this examination, because the potential of a descriptor can be better assessed in competition with other descriptors. Chemical fingerprint descriptors are powerful and well-explored instruments used to determine chemical similarity between chemical entities. It is common knowledge that three-dimensional virtual screening methods do not outperform less sophisticated chemical fingerprint descriptors.[21−23] Three-dimensional virtual screening methods are, however, expected to create intellectual property by connecting unexplored parts of the chemical space having biological activity to generate new biochemical space.[24] A powerful virtual screening method will detect active molecules with scaffolds that were so far unrelated to a biological target. Enrichment

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

Table 1. Sample of Target Names and Query Information for the Test Data Sets[a]

no.	target	# query molecules	ChEBI Id	MW	# heavy atoms	size pharmacophore	# subpharmacophores	# spike molecules
13	androgen receptor	3	412268	849	62	18	56448	137
20	carbonic anhydrase II	38	404394	600	40	14	7855	461
29	cysteinyl leukotriene receptor 1	24	131997	645	45	10	581	33
39	dopamine transporter	17	460879	564	41	11	1254	142
40	endothelin receptor ET-A	4	238806	610	44	15	14443	13
60	HERG	18	481922	817	57	11	1247	12
62	histamine H3 receptor	1	439626	479	35	5	1	30
76	melatonin receptor 1A	2	479903	592	43	12	2508	171
77	melatonin receptor 1B	2	479903	592	43	12	2508	189
84	neprilysin	1	106694	543	37	11	1254	11
87	neuronal acetylcholine receptor protein alpha-4 subunit	1	183398	554	38	8	92	156
96	phosphodiesterase 4A	8	322093	637	47	11	1254	21
101	progesterone receptor	2	412268	849	62	18	56448	52
110	serotonin 1a 5-HT1a receptor	19	558808	482	35	11	1254	438

[a]The complete table is available as Supporting Information. no.: index in the complete table. # query molecules: number of query molecules fulfilling the criteria. ChEBI Id: identifier of the query molecule selected to generate the sub-flexophore descriptors. MW: molecular weight of the selected query molecule. # heavy atoms: number of non-hydrogen atoms in the selected query molecule. size pharmacophore: number of pharmacophore points in selected query molecule. # sub-pharmacophores: number of generated sub-pharmacophores after applying the mentioned restrictions. # spike molecules: number of active molecules fulfilling the criteria to become a member of the test data set.

rates are not a crucial criterion in the daily business of virtual screening in drug discovery. The goal of virtual screening in medicinal chemistry programs is to deliver a molecule or, better yet, a series of molecules to be used as starting points to generate patentable chemical entities. In a recent virtual screening experiment, the Flexophore descriptor demonstrated its capabilities in scaffold hopping.[17] Is it possible to employ the Flexophore descriptor to find small active molecules starting from a large query molecule?

2. METHODS

2.1. Data Source. All molecules in this study were taken from the ChEMBL database, version 05. The ChEMBL database is a collection of information on bioactive compounds compiled from multiple sources.[19] It contains more than 500 000 structures with almost 3 million activity values. Only molecules containing only the following atoms were considered as a query, spike, or decoy: H, C, N, O, F, S, Cl, Br, and I. The aim of this study was to analyze the use subpharmacophore models from large query molecules to find small active molecules. Following, at least one large active structure needed to be known to add a target to the test data set. Two selection criteria were chosen for query molecules: a molecular weight between 450 and 850 Da—the upper threshold avoided too many possible subpharmacophore point combinations. Second, molecules with less than five pharmacophore points were excluded as a query. For each target, the molecule possessing the most pharmacophore points became the query. The subpharmacophore descriptors were derived from the Flexophore descriptor of this molecule. These subpharmacophore models were used to search the test data sets. Some query molecules show up for more than one target. A limiting condition for the creation of a test data set was a unique combination of query and spike molecules. Examples of targets in which the same query occurred more than once are the GABA receptors, the serotonins, and the group of phosphodiesterases.

2.2. Test Data Set Preparation. For spike molecules, the following criteria were set: ChEMBL standard activity value > 0 and ≤ 200 nmol, and being of "Ki" activity type. The molecular weight was limited to 200 ≤ spike or decoy ≤ 400. Nine was the minimum number of heavy atoms (non-hydrogen). The minimum number of spike molecules for a target was set to 10. To guarantee a minimum diversity between the spike molecules, the maximum chemical similarity between two spike molecules was set to a Tanimoto of 0.9, using the Actelion Fragment Fingerprint descriptor. One test data set consisted of 10 or less spike molecules and 990 decoys, which were selected randomly from the ChEMBL database. No selection criteria were set for the decoys, except for molecular weight restrictions and that the molecule may not be active on the target under consideration. As many test sets were created as necessary to use up all spike molecules. Consequently, some test data sets contained less than 10 spike molecules. After processing the complete ChEMBL database, 133 targets remained that fulfilled the criteria defined above. The spike molecules were distributed among 1079 test data sets (Table 1, complete tables are available in the Supporting Information). The largest data set was compiled for carbonic anhydrase II. In total, 38 molecules fulfilled the conditions as query molecules, and 461 spike molecules were extracted from the ChEMBL database. In all, 47 test data sets were generated from the spike molecules. For the last test set, only one spike molecule remained. The query molecules for the androgen and the progesterone receptor exhibited, with 18 pharmacophore points each, the largest pharmacophores. A total of 56 448 query descriptors were derived from each of the query molecules. For each test molecule, 56 448 similarity calculations had to be performed, with the highest value taken as a similarity score.

2.3. Descriptors. *Flexophore and Sub-Flexophore.* The generation of the Flexophore descriptor and its similarity metric were recently described in detail.[18] The Flexophore descriptor represents the molecule using a complete graph. This is a simple graph in which each pair of vertices is connected at an edge. The vertices are labeled with enhanced MM2 atoms types.[25] A vertex can contain several labels. The edges are histograms of the vertex distances, resulting from diverse molecule conformations (Figure 1). The similarity function

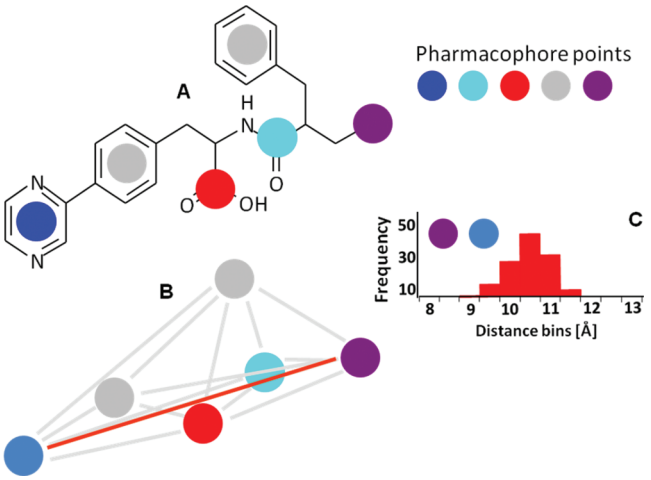


Figure 1. Flexophore descriptor. (A) Molecule with pharmacophore points, defined by substructures. (B) Complete graph. (C) One edge of the complete graph; the color of the edge is the distance histogram of the pair of connected pharmacophore points.

introduces the target point of view into the descriptor comparison. Interaction statistics for the enhanced MM2 atom types were derived from the crystallographic data in the Protein Data Bank (PDB).[26] A similarity matrix of these atom types, concerning their interaction behavior, was created. The generated table contains the similarity between any two vertices. Edge similarities are calculated from the fraction of overlap between two distance histograms. To calculate the similarity of two Flexophore descriptors, the maximum common substructure (MCS) of two descriptor graphs is determined, and an overall similarity score is calculated for the best match among all vertices and edges. A threshold for the node similarity is used to distinguish between the matching and nonmatching nodes. The comparison of two descriptors is computationally expensive. Since the graphs are complete, all pairs of nodes have to be compared. This means a full permutation of the nodes of the smaller graph over all nodes of the larger graph. However, the algorithm for a single substructure match is shortened by breaking off the comparison process after detecting a nonmatching node pair by using the node similarity threshold. The similarities for all following permutations containing this nonmatching node pair do not have to be calculated. A Java applet to explore the Flexophore descriptor is available at http://www.cheminformatics.ch in the submenu Tools→Flexophore. Sub-Flexophores are a subset of pharmacophore points from a Flexophore descriptor, together with all connecting nodes (Figure 2). The sub-Flexophore is a subgraph of a whole Flexophore but still a complete graph technically.

Query by Subpharmacophores. Since a Flexophore descriptor is defined as a complete graph, subpharmacophores can be generated by removing one or more nodes and their correlated edges. The generated subpharmacophore is a valid Flexophore descriptor and does not need any further treatment, which means that it contains the Flexophore descriptor, calculated from the query molecule for the androgen receptor, with 18 pharmacophore points. Preliminary experiments showed that at least five pharmacophore points in a subpharmacophore are needed to reach sufficient selectivity in virtual screening. The number of possible combinations is given by the binominal coefficient k over n, with k = number of pharmacophore points in the sub-Flexophore and n = number of pharmacophore points in the query Flexophore. There exist 8568 different sub-Flexophore descriptors with five pharmacophore points for molecules with 18 pharmacophore points (Table 2). For six pharmacophore points, 18 564 different sub-

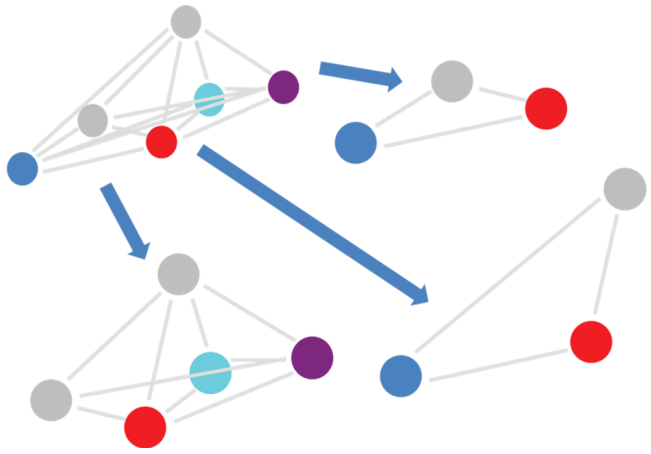


Figure 2. Sub-Flexophore descriptors derived from the Flexophore descriptor by extracting complete subgraphs.

Table 2. Combinatorial Explosion for Sub-Pharmacophores[a]

PPPoints	10	12	14	16	18	20
5	252	792	2002	4368	8568	15 504
6	210	924	3003	8008	18 564	38 760
7	120	792	3432	11 440	31 824	77 520
8	45	495	3003	12 870	43 758	125 970
9	10	220	2002	11 440	48 620	167 960
10	1	66	1001	8008	43 758	184 756
11		12	364	4368	31 824	167 960
12		1	91	1820	18 564	125 970
13			14	560	8568	77 520
14			1	120	3060	38 760
15				16	816	15 504
16				1	153	4845

[a]First row: number of pharmacophore points in query molecule. First column: number of pharmacophore points in sub-pharmacophore descriptor. Remaining fields contain the number of different sub-pharmacophores.

Flexophore descriptors can already be generated and 31 824 for seven pharmacophore points. This results in 58 956 sub-Flexophore descriptors in total. Consequently, all Flexophore descriptors in the test data set have to be compared with this large number of sub-Flexophore descriptors. The number of pharmacophore point combinations can, however, be reduced by some straightforward restrictions. Only Flexophore descriptors containing at least one nonaliphatic pharmacophore point were allowed. Nonaliphatic refers to any substructure containing at least one nitrogen or oxygen. The maximum number of pharmacophore points was set to seven. This threshold was derived by analyzing the relation between the number of pharmacophore points and the molecular weight in the ChEMBL database (Table 3 and Figure 3). There is a linear relationship between the number of pharmacophore points and the molecular mass (Figure 3). From Table 3, it can be seen that the median molecular weight for a molecule with seven pharmacophore points comes close to 400 Da, with almost 450

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

Table 3. Pharmacophore Point versus Molecular Weight Distribution in the ChEMBL Database[a]

	molecular weight			
# PPPoints	1. quartile	median	3. quartile	# molecules
3	184	227	282	14 791
4	227	274	327	37 202
5	**266**	**314**	**367**	**63 653**
6	**302**	**354**	**410**	**79 634**
7	**336**	**391**	**446**	**78 095**
8	367	428	488	64 316
9	397	462	528	45 574
10	425	497	571	28 761

[a]PPPoints: number of pharmacophore points in the Flexophore descriptor. The columns 1. quartile, median, and 3. quartile: molecular weight distribution of the molecules corresponding to the number of pharmacophore points. # molecules: number of available molecules in the ChEMBL database to calculate the weight distribution.

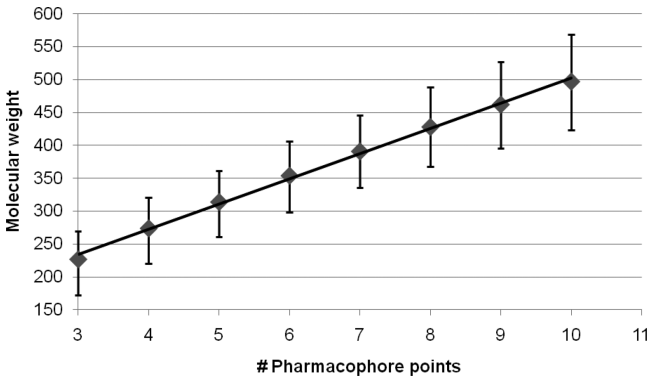


Figure 3. Pharmacophore point versus molecular weight distribution. Data were taken from ChEMBL database 05. The number of pharmacophore points and molecular weight were taken from Table 2. *x* axis: Number of pharmacophore points in Flexophore descriptors from molecules in the ChEMBL database. *y* axis: Median values of molecular weight distribution for the corresponding number of pharmacophore points. Error bars indicate the first and the third quartile of the molecular weight distribution.

Da being the upper limit of the third quartile. This exceeds the desirable molecular weight for a hit that could become the starting point for a medicinal chemistry program. The Flexophore descriptor similarity between any of the query subpharmacophores is limited to 0.9. Flexophore descriptor pairs showing a higher similarity are a redundant description of the covered pharmacophore space covered by the query Flexophore descriptors. Applying these restrictions to the query for the androgen receptor resulted in a descriptor set with 56 448 sub-Flexophore descriptors.

Chemical Fingerprint Descriptors. Three chemical fingerprint descriptors were used in this study for comparison purposes. All three chemical fingerprint descriptors are topological descriptors and are therefore devoid of 3D information. The chemical fingerprint analogue to a sub-Flexophore was generated by extracting all atoms contributing to the pharmacophore points in the sub-Flexophore and all atoms needed to link the contributing atoms. Linking atoms were defined as those atoms on the shortest path between two pharmacophore points. The result is a structural fragment from which the chemical fingerprint descriptor is calculated, as from any common molecular structure.

Actelion Fragment Fingerprint (FragFp). FragFp is a dictionary-based descriptor with a length of 512 bits. Each bit represents a substructure fragment, of which some include wildcard atoms.[27] The dictionary of 512 substructures was created by a computational procedure which had been optimized to balance two goals: (1) any one of these fragments should occur frequently in organic molecule structures, and (2) all fragments should be linearly independent concerning their substructure match pattern in diverse organic compound sets. To generate a descriptor vector, the molecular structure is searched for any of the substructures in the dictionary. For any match, the corresponding bit of the vector is set to 1. Any molecular structure is represented by a binary vector of length 512. The FragFp descriptor belongs to the same class as the "MDL structure keys,"[28] which have recently been shown to perform better in virtual screening than 3D descriptors.[23]

Actelion Path Fingerprint (PathFp). PathFp is a molecular graph-path, walking fingerprint descriptor. All distinguishable paths with up to seven atoms are hashed into a descriptor vector of 512 bits. This descriptor is conceptually similar to ChemAxCFp, the chemical fingerprints from ChemAxon,[29] and to the Daylight[30] fingerprints.

Actelion Sphere Fingerprint (SphereFp). SphereFp locates three circular layers with increasing bond distance around each atom. This yields four fragments, starting with the naked central atom and adding one layer at a time. Every fragment is encoded as a canonical string (ID code), similar to the generation of canonical SMILES.[31,32] The string is then assigned to one of 1024 unique bits. For this, the hash value of the ID code is calculated, and the corresponding bit in the vector is set. The Hashlittle algorithm developed by Jenkins[33] is used for its binning function, which takes a text string as input and returns an integer value between 0 (inclusive) and 1024 (exclusive). In preliminary experiments, this hash function exhibited a good, uniform distribution of the generated hash values. A calculation of $e = 4$ spheres is done for all atoms g in the molecule serving once as the center atom, which results in $b = eg$ number of ID codes. The number of bits set in the descriptor vector is $\leq b$, depending on the number of unique ID codes and hash collisions.

2.4. Similarity and Enrichment Rate Calculations. For a similarity calculation between a query molecule and a test-set molecule, all subdescriptors, derived from the query molecule, have to be compared with the descriptor from the test molecule. The similarity value from the best matching descriptor pair is taken as the score. After calculating all scores for a test data set, the results are sorted according their scores in descending order. For the following enrichment rate calculation, only molecules with a minimum similarity score of 0.85 were considered. This is, based on results from preliminary experiments, a reasonable cutoff value for the Flexophore descriptor as well as for the chemical fingerprint descriptors used in this study. The experience was made that a molecule showing a similarity below this value to a query molecule is too dissimilar to be considered for biological screening. Therefore, this cutoff value prevents an overoptimistic outcome of the enrichment rate. This is because, without this cutoff value, a low similarity for the spike molecules would already be sufficient, assuming the similarity of the decoys is lower. As a figure of merit, the relative enrichment rate e_{rel} at 1.0% of the total data set was calculated.[17] Due to the low enrichment rates of the virtual screening, for simplification purposes, the result was rounded to a binary value. Thus, if the enrichment rate for a

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

Table 4. Subsample of the Result Table for the Virtual Screening Experiments[a]

no.	target	data sets	FragFp	SphereFp	PathFp	Flexophore	LE query	LE median Flexo
13	androgen receptor	14	2	0	1	1	0.161	0.363
20	carbonic anhydrase II	47	0	0	1	2	0.269	0.435
29	cysteinyl leukotriene receptor 1	4	0	0	0	2	0.271	0.377
39	dopamine transporter	14	1	0	1	14	0.248	0.424
40	endothelin receptor ET-A	2	0	0	0	0	0.231	
60	HERG	2	0	0	0	0	0.203	
62	histamine H3 receptor	3	0	1	2	1	0.334	0.407
76	melatonin receptor 1A	18	3	1	2	14	0.256	0.594
77	melatonin receptor 1B	19	3	1	2	17	0.25	0.61
84	neprilysin	2	0	0	0	0	0.311	
87	neuronal acetylcholine receptor protein alpha-4 subunit	15	2	0	1	2	0.277	0.586
96	phosphodiesterase 4A	3	1	0	1	2	0.225	0.464
101	progesterone receptor	6	2	0	0	0	0.192	
110	serotonin 1a receptor	43	3	0	0	14	0.263	0.393
	sum of data sets	**1079**	**187**	**84**	**144**	**262**		
	enriched targets		**72**	**36**	**67**	**85**		
	unique hits		**35**	**2**	**22**	**185**		

[a]Column three contains the number of test data sets. The four columns that follow indicate in how many data sets sets enrichment for the named descriptor was observed. FragFp, SphereFp, and PathFp are chemical fingerprint descriptors. Flexophore is the 3D sub-pharmacophore descriptor. LE query is the ligand efficiency of the query molecule, from which the sub-Flexophores were derived. LE median Flexo is the median ligand efficiency of the spike molecules enriched by virtual screening with the Flexophore descriptor. The number of values used for the median is equal to the number in the "Flexophore" column. The summary in the last three rows is for the complete table and is available as Supporting Information.

data set with its 1000 molecules was larger than zero, a 1 was added to the results table (Table 4). The original enrichment rates are available as Supporting Information.

Ligand efficiency calculation. Ligand efficiency was calculated according to Hopkins et al.[6] The exact formula used was $le = -1.986 \times 0.3 \ln(0.000000001k)/n_{\text{heavyAts}}$, with k being the binding constant from the standard value field in the ChEMBL database in nmol and n_{heavyAts} being the number of non-hydrogen atoms. The factor 0.000000001 standardizes the K_i values in the ChEMBL database from nmol to mol. With many heavy atoms and a high binding constant, the ligand efficiency becomes almost 0. With a decreasing number of atoms and a decreasing binding constant, the ligand efficiency increases. In Table 4, the ligand efficiency for all query molecules is given. The ligand efficiency of the spike molecules enriched by the Flexophore descriptor was summarized by their median.

3. RESULTS

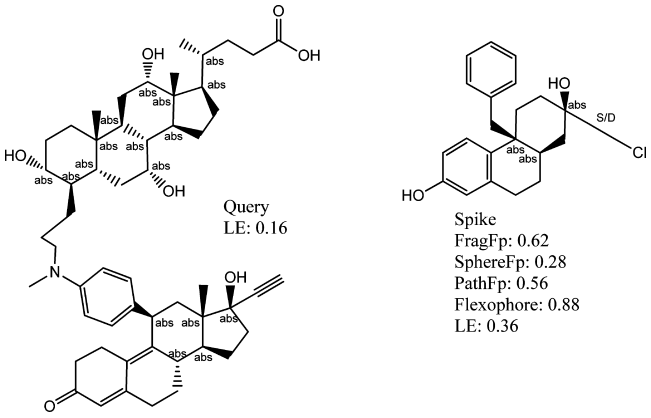
In total, 371 077 subpharmacophores were compared with 1 079 000 test descriptors, resulting in approximately 40 billion similarity calculations. The results of the virtual-screening experiments on 133 targets with 1079 test data sets are summarized in Table 4. After the index, the target name and the number of data sets for each target are given. The following four columns indicate the number of test sets where enrichment was observed. In the column headed "LE query," the ligand efficiency values for the query molecules are given. The next column shows the median ligand efficiency of the test molecules enriched by the Flexophore descriptor. In the third-to-last table row, the sums of the data sets where enrichment was observed are given. The Flexophore descriptor found at least one spike molecule in 262 data sets out of 1079, followed by FragFp (187), PathFp (144), and SphereFp at the end with 84 enriched data sets. In the second-to-last table row, the sums of targets for which enrichment was observed are given. From this, it can be understood that the SphereFp descriptor detected at least one spike molecule for 36 targets out of 133. The PathFp and FragFp descriptors performed almost equally in detecting spikes for 67 and 72 targets, respectively. By detecting 85 spike molecules, the Flexophore descriptor performed best in this comparison. Analyzing the uniqueness of hits found, Flexophore identified 185 unique hits, FragFp, 35; PathFp, 22; and SphereFp, 2. In summary, the sub-Flexophore descriptor outperformed the three chemical fingerprints in terms of pure enrichment as well as in detecting unique molecules. The sub-Flexophore descriptor was most successful on the melatonin receptor 1B where at least one spike molecule was found in 17 out of 19 data sets, whereas the chemical fingerprint descriptors found hits in three data sets maximum. Similar behaviors were observed for serotonin 1a, melatonin 1A, and the dopamine transporter where the Flexophore succeeded on all 14 data sets. Carbonic anhydrase II was the protein with the largest number of test data sets (47). However, only the Flexophore descriptor and the PathFp descriptor succeeded on two and one data set, respectively. Considering the results for the ligand efficiency calculations, it is striking that in all cases the median ligand efficiency of the enriched molecules is higher than the ligand efficiency of the query molecules. This correlates with the medicinal chemistry experiment, described in the Introduction, where the ligands found by the sub-Flexophore descriptor exhibited also better ligand efficiency than the larger query molecule.

3.1. Single Results Discussion.

Androgen Receptor. The largest query molecule with a molecular weight of 849 and 62 heavy atoms was chosen for the androgen and the progesterone receptor (Chart 1), respectively. In the query molecule, two steroid scaffolds are linked via aniline and n-propyl. The steroid scaffolds are decorated with four hydroxyl groups, a carboxyl group, a carbonyl group, and an ethinyl group. All of these substructures were described by the Flexophore descriptor with 18 pharmacophore points. The possible number of pharmacophore point combinations for query subpharmacophores with five, six, and seven pharmaco-

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

Chart 1. Ligands for the Androgen Receptor[a]



[a]Query molecule and the spike molecule that was found by the sub-Flexophore descriptor. The values that follow the descriptor names indicate similarity to the best-matching query sub-Flexophore.

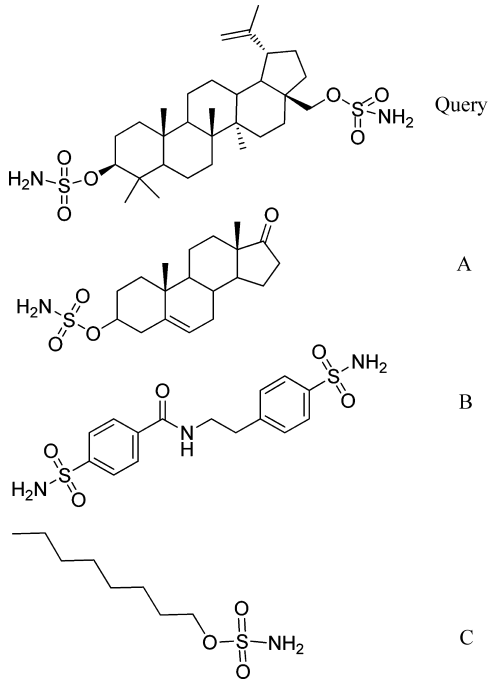
phore points would have been 58 956 but was shrunk by the side conditions to 56 448. This vast number of query descriptors resulted in the detection of only one spike molecule out of 137, which were distributed in 14 test data sets. Even this single hit, however, was not detected by any of the chemical fingerprint descriptors. For this hit with a molecular weight of 367, a standard activity of 130 nmol is given in the ChEMBL database. The ligand efficiency is, at 0.363, much better than for the query molecule. With 53 nmol, the concentration is 2 times lower for the query molecule than for the spike molecule, but the high molecular weight of the query reduces its ligand efficiency to 0.161.

Carbonic Anhydrase II. The test data sets for the carbonic anhydrase II contain 461 spike molecules from which only two were found by the Flexophore. The best chemical fingerprint was the PathFp by detecting one spike molecule. Chart 2 shows the query molecule which is a steroid with an additional, six-membered ring and two sulfonamides at both ends. Molecules A and B were found by the sub-Flexophore descriptors and molecule C by the PathFp descriptor. Molecule A is almost a substructural fragment of the query molecule. None of the chemical fingerprints was able to find it, however, because of the missing additional six-membered ring. In molecule B, scaffold hopping was observed; the steroid-like scaffold of the query molecules was replaced by two phenyl moieties linked by a simple carbon chain. An interesting hit was detected by PathFp, a carbon chain attached to the sulfonamide. It is the only hit detected by the PathFp descriptor for this target, but the similarity to the subquery descriptor is 1.0.

Cysteinyl Leukotriene Receptor 1. Partial scaffold hopping was possible for the cysteinyl leukotriene receptor 1 (Chart 3). A molecule with a molecular weight of 650 and 10 pharmacophore points served as the query. From these 10 pharmacophore points, 581 subpharmacophores were derived to be used as query descriptors. Three spike molecules out of 33 spike molecules were found. The detected spike molecules contain a naphthyl moiety like the query molecule, but the sulfonamide group is exchanged with a tetrazolium group.

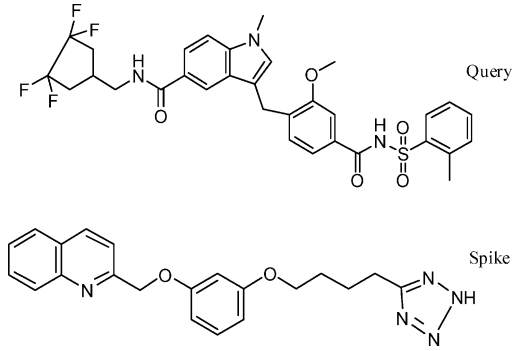
hERG Ion Channel. Subpharmacophore screening for the hERG ion channel failed; no spike molecule was found. Predicting activity on the hERG ion channel would be quite beneficial but is known to be difficult.[34] The query molecule has a molecular weight of 817 and is described with 11

Chart 2. Ligands for Carbonic Anhydrase II[a]



[a]Query and spike molecules. A and B found by sub-Flexophore descriptors. C found by PathFp (chemical fingerprint descriptor), similarity 1.0.

Chart 3. Ligands for the Cysteinyl Leukotriene receptor 1[a]



[a]Query and spike molecules found by sub-Flexophore descriptor.

pharmacophore points by the Flexophore descriptor. The hERG ion channel is well explored; nevertheless, only 12 spike molecules matching the previously defined criteria were compiled from the ChEMBL database. Two test data sets were compiled from these spikes: one with 10 and the other with two active molecules. Both data sets were filled up to 1000 molecules with decoys. The similarity scores were calculated to be between the pharmacophore descriptors of these 2000 molecules and the 1247 subpharmacophores derived from the query molecule. Only two decoys had a Flexophore similarity above the threshold of 0.85. The best spike molecule showed a Flexophore similarity of 0.75, which was equivalent to rank 12 in the test data set.

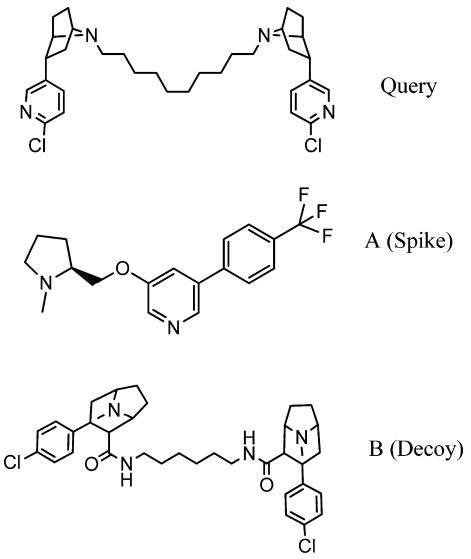
Histamine H3 Receptor. With five pharmacophore points and a molecular weight of 479, the smallest query molecule was chosen for histamine H3. No subpharmacophores were derived because this is already the minimum number of pharmacophore points for a subpharmacophore descriptor. This case is

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

equivalent to classical ligand-based virtual screening, where the complete ligand is used to calculate the descriptor. Nevertheless, only one spike molecule was found from the fragment descriptor, as well as from SphereFp and PathFp.

Neuronal Acetylcholine Receptor. In this example, the difference in molecular weight is less interesting than the difference in size (Chart 4). In the query molecule, a maximum topological distance of 27 bonds is measured between the two chlorine atoms, whereas the maximum topological distance in the spike molecule spans only 13 bonds. The differences in topology are also reflected by low similarity values (<0.5) for the three chemical fingerprint descriptors. However, the Flexophore descriptor mapped three pharmacophore points of the spike molecule onto three pharmacophore points of the query molecule (Figure 4), neglecting the large topological distance between the query pharmacophore points A1 and A2. This mapping became possible through the high flexibility of the carbon atom chain connecting A1 and A2. One possible solution for the mapping was calculated with a MM2 force field and is shown in Figure 5. Such a conformation, with a close distance between A1 and A2, was also generated by the conformation sampler during the Flexophore descriptor generation and stored in the distance histograms.

Chart 4. Ligands and Decoy for the Neuronal Acetylcholine Receptor 1, Query and the First Two Hits Found by Sub-Flexophores[a]



[a]A has a Flexophore similarity of 0.9 to the query molecule. B is a decoy molecule.

Neprilysin. This test data set is an example of a case in which the sub-Flexophore approach failed. The query molecule is phosphoramidon, a natural product derived from cultures of *Streptomyces tanashiensis* (Chart 5).[35] Eleven pharmacophore points were needed by the Flexophore descriptor to describe the query molecule. A total of 1254 sub-Flexophore descriptors were derived from the complete query Flexophore descriptor. The total of 1254 is also the maximum possible number of combinations for 5 out of 11, plus 6 out of 11, plus 7 out of 11, which means that all pharmacophore-point combinations with five, six, and seven pharmacophore points were accepted by the query generator. No sub-Flexophore was removed from the query Flexophore descriptor set because no similarity value for all pairs of subpharmacophores was above the similarity threshold of 0.9. The molecule contains only two pure aliphatic pharmacophore points and nine with hetero atoms. No five-to-seven pharmacophore-point combination is possible which contains only aliphatic pharmacophore points. This is the reason why also no subpharmacophores were removed, which is in accordance with the rule that solely aliphatic subpharmacophores are not allowed. The two test data sets contained 11 spike molecules. None of them was found by any of the descriptors. A similarity value of 0.75 was calculated as the best score for a spike (molecule 2 in Chart 5), which is equivalent to rank 12 in the test data set. The spike molecule contains only four pharmacophore points, and the similarity function of the Flexophore is not able to fit them properly enough onto one of the subpharmacophore query descriptors.

Phosphodiesterase 4A. A query molecule with a molecular weight of 637 Da and 11 pharmacophore points was fractionized into 1254 subpharmacophore points (Chart 6). Two spike molecules (A and B) out of 11 were detected by the sub-Flexophores. Both spike molecules exhibit structural features that resemble the query molecule. Yet only molecule

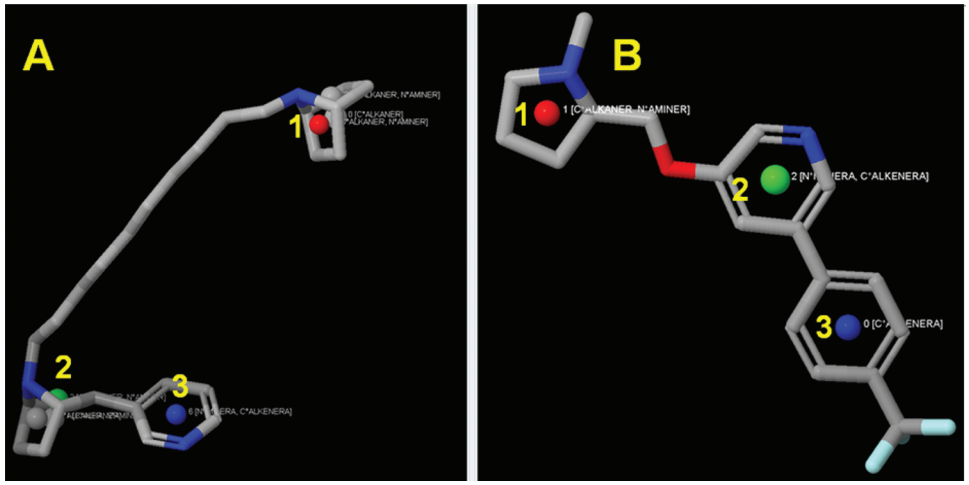


Figure 4. Neuronal acetylcholine receptor 1. Flexophore pharmacophore point mapping of the first hit A (active spike) from Chart 1 on the query molecule B. Mapping pharmacophore points are indicated by colored balls; nonmapping pharmacophore points are indicated by gray balls.

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

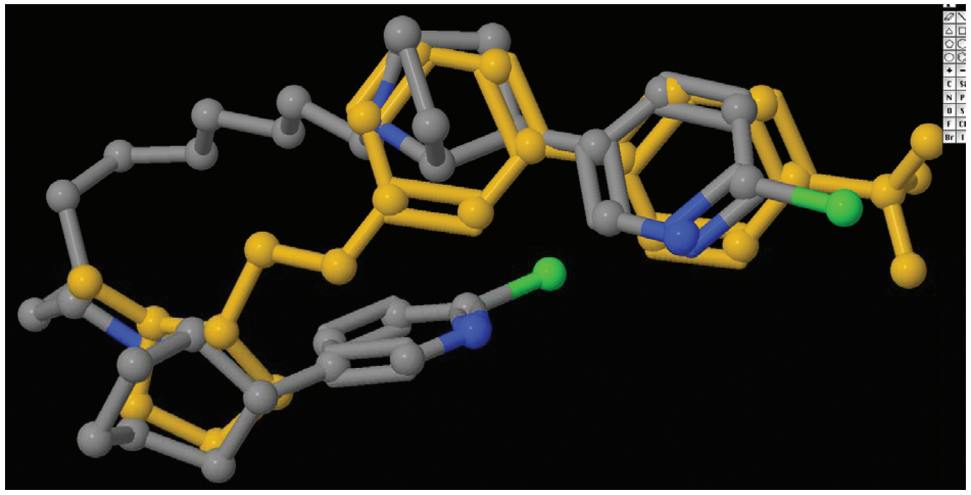
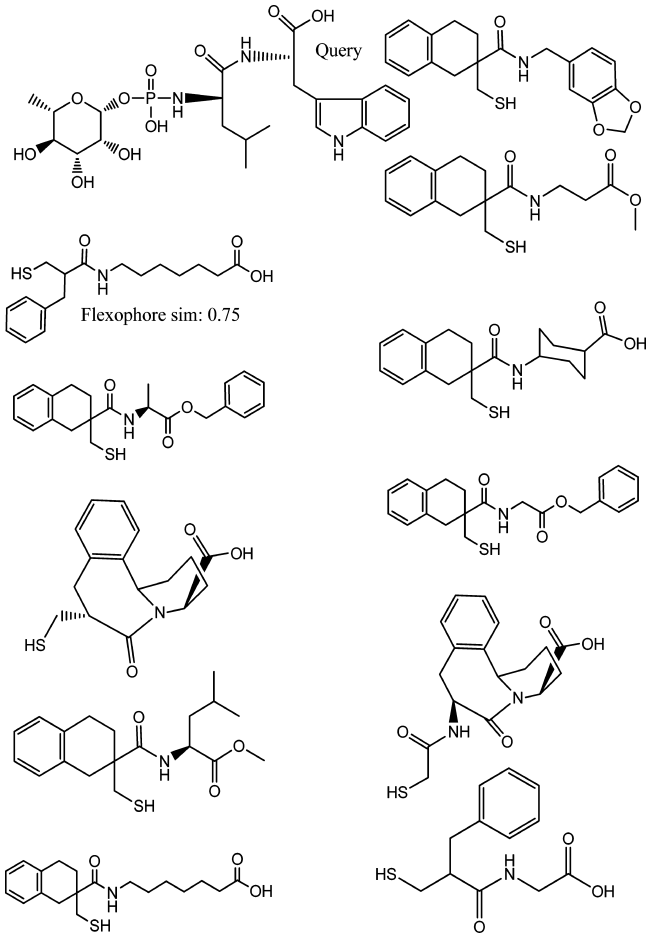


Figure 5. Neuronal acetylcholine receptor 1. Superpositioning of the query and spike molecules (Chart 4). Hit molecule in yellow, used force field: MM2.

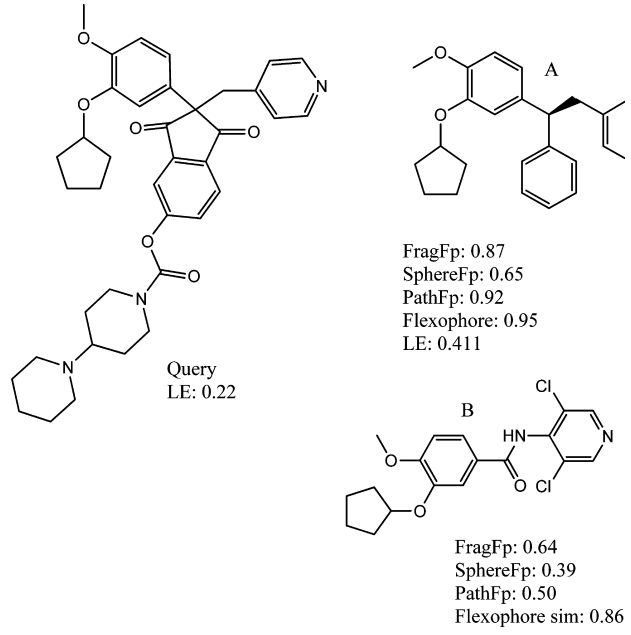
Chart 5. Ligands for Neprilysin, Query, and All Spike Molecules[a]



[a]No spike molecule was found by any of the descriptors. The best Flexophore similarity was 0.75.

A was also detected by two of the chemical fingerprints. Molecule B was only found by the sub-Flexophore descriptor. In molecule B, one of the two carbonyl groups was mimicked by the amide function. Together with the two chlorine atoms,

Chart 6. Ligands for Phosphodiesterase[a]



[a]Query molecule and two spike molecules. The values that follow the descriptor names indicate the similarity to the best-matching query sub-Flexophore. LE: ligand efficiency.

these changes were already sufficient to mislead the chemical fingerprint descriptors.

4. SUMMARY AND CONCLUSIONS

The goal of this study was to test the hypothesis that low-weight bioactive molecules can be found with an obese query molecule. More than 1000 test data sets for 133 targets were compiled from the ChEMBL database. Molecules with a molecular weight of less than or equal to 400 were used to spike the test data sets. Active molecules with a molecular weight of higher than or equal to 450 were used to derive the query descriptors. A linear relationship was observed between the number of pharmacophore points and the molecular mass by analyzing the ChEMBL database. This relation indicates that compounds with fewer than eight pharmacophore points are

dx.doi.org/10.1021/ci200402r | J. Chem. Inf. Model. 2012, 52, 380−390

desirable as lead compounds to stay below a molecular weight of 400. After building a test data set considering this and other restrictions, the Flexophore succeeded as the best descriptor on 85 out of 133 targets and in 262 out of 1079 test-data sets. If a hit was found, it was a singleton in the majority of cases. This is a much lower hit rate than reported in numerous reports about virtual screening. One reason is the difference in size of the query and spike molecules. Additionally, there is not always evidence that the query and spike molecules bind to the same region in the target. If this is the case, any virtual screening approach will fail. An implicit feature of the Flexophore descriptor is that it does not take into account the space-filling shape of the underlying molecule. Pharmacophore points in the descriptor are related by defined distances, but there is no physical shape representation in the 3D space. On the other hand, the virtual screening experiment for the histamine H3 receptor indicates that the hit rate for descriptors calculated from complete molecules is not much higher, neither for the Flexophore descriptor nor for the chemical fingerprint descriptors. That the Flexophore descriptor and the three chemical fingerprint descriptors perform quite well in traditional ligand-based screening was already demonstrated in a recent publication.[17] In conclusion, the reason for the low performance of the descriptors in terms of enrichment rates is the structure of the data sets used here. Only molecules from the ChEMBL data set were taken for test data set generation. This means that only molecules which were biologically tested and, in the majority of cases, found to be biologically active were included. In contrary to subpharmacophore-based virtual screening, classical ligand-based virtual screening does not aim to find significant, smaller molecules. For this purpose, the subpharmacophore screening was a success. Hits were found for the majority of targets, and all exhibited higher ligand efficiency than the query molecule. Scaffold hopping was observed in several cases; a detailed discussion was given for cysteinyl leukotriene, neuronal acetylcholine, and phosphodiesterase. To jump from a large molecule to a smaller one with improved ligand efficiency is the real benefit of the subpharmacophore approach. Taking together the results from the test data sets and the new series of bioactive molecules retrieved from the application of subpharmacophore screening in the medicinal chemistry project, which was mentioned in the Introduction, subpharmacophore screening has proven to be a valuable addition to the toolbox of computational chemistry.

5. COMPUTATIONAL DETAILS

As not otherwise mentioned, all algorithms were implemented by the authors using Java. The similarity calculations took five days on a common workstation with two processors.

■ ASSOCIATED CONTENT

Supporting Information

Table with ChEBI IDs for the query molecules. SD file with query structures. Table with detailed results of the virtual screening experiments. Table with summarized results of the virtual screening experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: + 41 61 565 63 23. Fax: +41 61 565 65 00. E-mail: modest.korff@actelion.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Christian Rufener and Roman Bär implemented Actelion's computing grid. Christian Rufener also developed the PathFp. We thank Russell Jones for editorial assistance.

■ REFERENCES

(1) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

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

(3) Lieb, W. R.; Stein, W. D. Biological membranes behave as nonporous polymeric sheets with respect to the diffusion of nonelectrolytes. *Nature* **1969**, *224*, 240−243.

(4) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011**, *2*, 349−355.

(5) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.

(6) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430−431.

(7) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(8) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.

(9) Gund, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117−143.

(10) Sheridan, R. P.; Nilakantan, R.; Rusinko, A.; Bauman, I. N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255−260.

(11) van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225−251.

(12) 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.

(13) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(14) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach. *J. Mol. Struct.* **2000**, *503*, 17−30.

(15) Langer, T.; Wolber, G. Feature-based Pharmacophores: Virtual Screening for Lead Identification. *G.I.T. Lab. J., Eur.* **2003**, *7*, 208−210.

(16) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorical Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(17) Korff, M. v.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 209−231.

(18) Korff, M. v.; Freyss, J.; Sander, T. Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *J. Chem. Inf. Model.* **2008**, *48*, 797−810.
(19) Brooksbank, C.; Cameron, G.; Thornton, J. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* **2009**, *38*, D17−25.
(20) Bender, A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309.
(21) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(22) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.
(23) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802−6810.
(24) Korff, M. v.; Hilpert, K. Assessing the Predictive Power of Unsupervised Visualization Techniques to Improve the Identification of GPCR-Focused Compound Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 1580−1587.
(25) Allinger, N. L.; Zhou, X.; Bergsma, J. Molecular Mechanics Parameters. *J. Mol. Struct.* **1994**, *312*, 69−83.
(26) 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.
(27) Sander, T. *ActelionFp*; Department of Research Informatics, Actelion: Allschwil, Switzerland, 2002.
(28) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(29) ChemAxon. GenerateMD. http://www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed November 8, 2007).
(30) Daylight Fingerprints. http://www.daylight.com/meetings/summerschool98/course/basics/fp.html (accessed July 30, 2008).
(31) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(32) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(33) Jenkins, R. Hash Functions and Block Ciphers. http://burtleburtle.net/bob/c/lookup3.c (accessed April 11, 2008).
(34) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83−92.
(35) Kitagishi, K.; Hiromi, K. Binding between thermolysin and its specific inhibitor, phosphoramidon. *J. Biochem.* **1984**, *95*, 529−534.