Comparison of Three Preprocessing Filters Efficiency in Virtual Screening: Identification of New Putative LXRβ Regulators As a Test Case

Léo Ghemtio,* Marie-Dominique Devignes, Malika Smaïl-Tabbone, Michel Souchet,† Vincent Leroux, and Bernard Maigret*

Nancy Université, LORIA, Groupe ORPAILLEUR, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France

Received September 18, 2009

In silico screening methodologies are widely recognized as efficient approaches in early steps of drug discovery. However, in the virtual high-throughput screening (VHTS) context, where hit compounds are searched among millions of candidates, three-dimensional comparison techniques and knowledge discovery from databases should offer a better efficiency to finding novel drug leads than those of computationally expensive molecular dockings. Therefore, the present study aims at developing a filtering methodology to efficiently eliminate unsuitable compounds in VHTS process. Several filters are evaluated in this paper. The first two are structure-based and rely on either geometrical docking or pharmacophore depiction. The third filter is ligand-based and uses knowledge-based and fingerprint similarity techniques. These filtering methods were tested with the Liver X Receptor (LXR) as a target of therapeutic interest, as LXR is a key regulator in maintaining cholesterol homeostasis. The results show that the three considered filters are complementary so that their combination should generate consistent compound lists of potential hits.

INTRODUCTION

Specific protein−substrate binding interactions are central to many biological processes and pathways and are, therefore, important targets for drug design. The quest for new drugs aimed at modulating specific biomolecular interactions relies on many approaches that are constantly being improved. These approaches that are part of the early phases of drug discovery are particularly critical.[1] In the past decade, high-throughput screening (HTS) technologies have used robotics and miniaturized assays to automate the identification of proprietary lead molecules. HTS techniques can lead to clear success, but because they require expensive lab equipment, their cost is a major disadvantage. The use of virtual high-throughput screening (VHTS) methods is an alternative that can provide reduced cost results as a complement to HTS.[2,3] However, computational requirements are important limiting factors for productive large VHTS campaigns even if advances in CPU speed, increased parallelism, and distributed computing (grid computing) promise to reduce run times. Indeed, in spite of such improvements, when huge chemical libraries[3−5] (>1 million compounds) are screened, VHTS programs based on classical docking algorithms still remain slow and require huge computer facilities. Another difficulty with classical docking programs is related to the significant number of false positives that will be produced. This mainly comes from the definition of the scoring functions which are used, and several studies have been designed to address these issues (evaluation of docking algorithms and scoring functions).[6−12]

Two main questions can be raised: (i) Taking advantage of the chemical diversity present in huge libraries containing billions of compounds, how is it possible to reduce the number of molecules that will be submitted to classical docking procedures? (ii) Can this be carried out at a reasonable computing cost without any loss in the accuracy of the prediction? One possible answer to both questions is to restrict the number of molecules to be docked by filtering techniques. Such multistep strategies have already been described and applied to the identification of putative hits.[7,12−17] The main goal of filters aims at removing compounds that do not have proper criteria qualifying them as possible "hits" from the source chemical libraries. The filters can be target-independent (such as Lipinski rules, nontoxicity constraint, etc.) or target-dependent. Among the possible prefiltering approaches, the use of shape descriptors is known to work extremely fast, and its usefulness has been confirmed in several papers.[14,18−23] Nevertheless, several other filters can be proposed to complement the VHTS process, such as those based on pharmacophore analysis of available three-dimensional (3D) structures of protein/ligand complexes, on chemical fingerprint, and on knowledge-based approaches.[24−33]

In this paper, we investigate three filtering strategies with regard to their respective speed and accuracy in a VHTS context. The first one is a fast shape-based preliminary rigid docking, which is orders of magnitude faster than classical docking methods and has already been reported as a suitable approach to hit identification.[18,22,23,34−38] The second one is a 3D pharmacophore-driven selection.[33,39,40] The third one is a knowledge-based approach called knowledge discovery from database (KDD).[41,42] Several reasons motivate the choice of the present filtering strategies. A shape-matching algorithm based on the use of spherical harmonics was

* Corresponding author e-mail: leo.ghemtio@loria.fr (L.G.); bernard.maigret@loria.fr (B.M.).
† Present address: Harmonic Pharma, 615 rue du Jardin Botanique, 54600 Villers les Nancy, France.

10.1021/ci900356m
Published on Web 04/26/2010

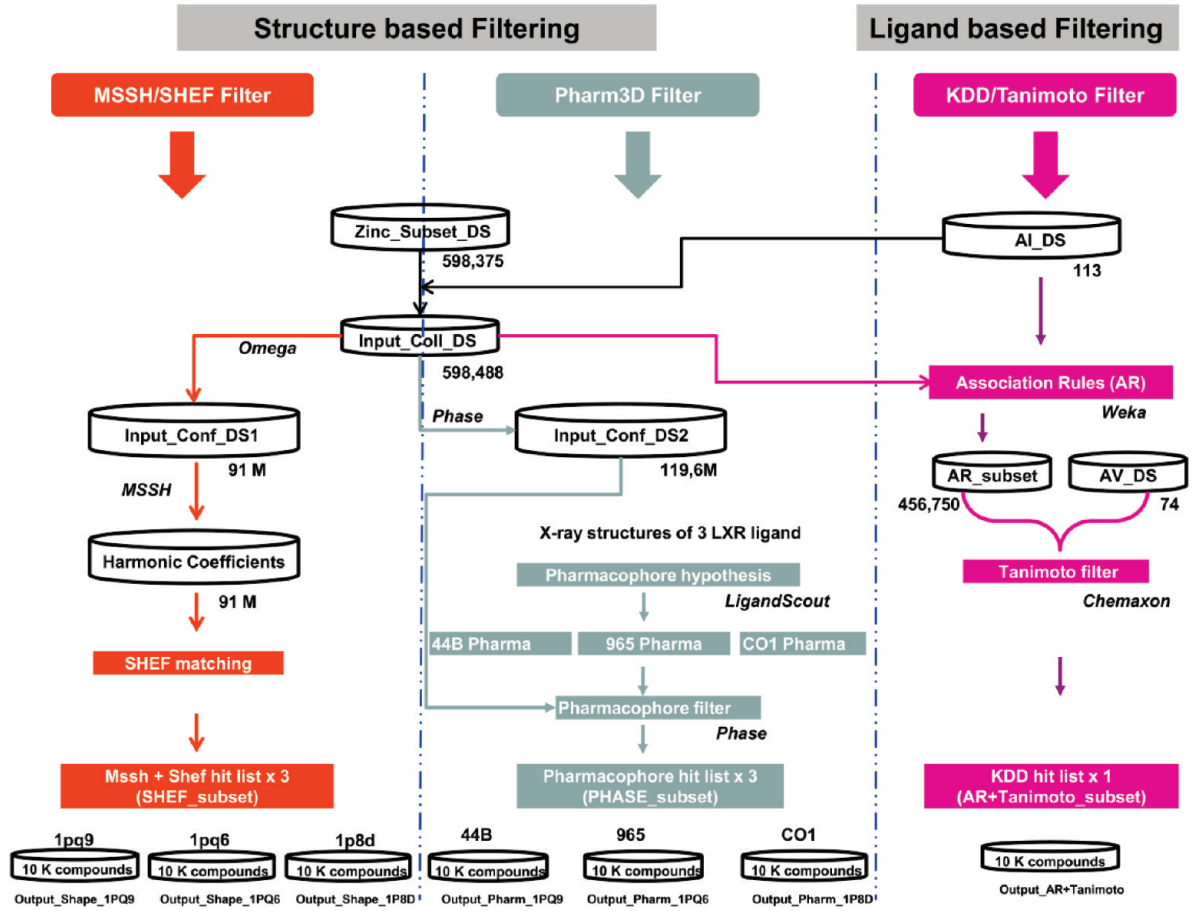

Figure 1. General flowchart of the three filtering strategies. There are three main compound data sets (Zinc_Subset_DS, AI_DS, and Input_Coll_DS, see Table 1), which are used for the implementation of the three considered filters. At the end, we have 3 × 10 000 compounds selected by the MSSH/SHEF shape filter, 3 × 10 000 compounds selected by the Pharm3D filter (for each of these filters the three crystallographic structures of the target can be separated), and 10 000 compounds selected by the KDD/Tanimoto filter.

already developed in our group and implemented in our "virtual screening manager for computational grids" (VSM-G) platform.[43] The use of pharmacophores is a technique which is now well documented in the literature, and which is going to be popular in virtual screening.[24,26,29,33,39,44,45] KDD approaches aim to identify valid, novel, useful, and understandable knowledge units from large data sets. Data mining (DM) is the core of the KDD process and involves algorithms that explore data, develop models, and discover significant patterns that can be turned into knowledge units. Thus a KDD filter could reveal a new methodology for VHTS by applying the discovery of association rules or other machine learning models to physicochemical descriptors of known interacting ligands.[42,46]

All three filters were applied to a chemical library containing drug-like compounds from a collection of chemicals. The methodology was tested against the nuclear hormone receptor Liver X Receptor β (LXRβ).[47,48] LXRβ is of therapeutic interest especially in metabolic disorders. Several X-ray 3D structures are available, providing insight into the protein binding site flexibility, which has been studied in detail and reported in a previous VHTS study.[43,49–51] Results obtained with the three filtering methods are analyzed and compared in terms of performance and quality of the resulting compounds. Criteria, such as chemical diversity and drug- and lead-likeness are considered in order to compare the respective selection of compounds.

METHODS

The general flowchart of the methodology is presented in Figure 1 and indicates the nature of input and output data of each filtering strategy.

Data Preparation. *Selection of LXRβ Crystal Structures.* This work specifically considers the β isoform of Liver X receptors, since several X-ray structures are available in the Protein Data Bank (PDB)[52] namely the 1p8d,[51] 1pq6,[50] and 1pq9[50] entries. The differences between these structures reveal a large plasticity of the ligand binding pocket to accommodate compounds with noticeably different shapes and sizes.[50] For each X-ray structure, the most complete chain was retained (chain A for 1p8d and chain B for 1pq6 and 1pq9) and eventually completed in order to be used in docking calculation (this preparative procedure has already been described).[43]

Preparation of Ligand Data Set. The ligands used in this study were collected from the Zinc[5] database and from various structure activity relationship (SAR) studies reported in the literature.[53–55] Table 1 summarizes the various compound data sets prepared for the purpose of this study.

Zinc Subset Data Set (Zinc_Subset_DS): This data set (called here Zinc_Subset_DS) is composed of molecules present in the Zinc database and commercially available from three chemical suppliers, ChemDiv,[56] Enamine,[57] and Comgenex.[58] A prefiltering procedure using Lipinski rule-of-five[59]

Table 1. Description of the Different Data Sets Used in This Study (M = Million)

name	population	content and usage	source
Zinc_Subset_DS	598 375	chemical compound library	commercial molecules from the Zinc DS (March, 2006)
AI_DS	113	library composed of 70 active and 43 inactive molecules for KDD learning	LXR SAR studies reported in the literature[53–55]
AV_DS	74	a second library composed of active molecules for ranking the compounds in the KDD filter	LXR SAR studies reported in the literature[53–55]
Input_Coll_DS	598 488	union of Zinc_Subset_DS and AI_DS; main compound library	see above
Input_Conf_DS1	92 M	conformer library for the MSSH/SHEF shape filter	generated by the Omega program from Input_Coll_DS
Input_Conf_DS2	119.6 M	conformer library for the Pharm3D filter	generated by the phase program from Input_Coll_DS
AR_subset	456 745	compounds predicted as active by association rules extracted from AI_DS	Input_Coll_DS and KDD filter
SHEF_subset (Output_Shape_1p8d, Output_Shape_1pq6, Output_Shape_1pq9)	30 000	compounds which have good shape complementarity score with the three known LXR structures	Input_Conf_DS1 and shape filter
PHASE_subset (Output_Pharm_1p8d, Output_Pharm_1pq6, Output_Pharm_1pq9)	30 000	compounds selected by the phase program as corresponding to the 3D pharmacophore definition of three known LXR ligand structures	input_Conf_DS1 and 3D pharmacophore filter
AR+Tanimoto_subset (Output_AR+Tanimoto)	10 000	compounds from AR_subset ranked by similarity with known active AV_DS compounds	ranking of the AR_subset compounds according to their Tanimoto scores with the active compounds of AV_DS

was performed, allowing a single violation for each structure, giving rise to a total of 598 375 unique molecules.

Active and Inactive Compound Data Set (AI_DS): This data set is composed of 113 active and inactive molecules randomly chosen among all the compounds considered in SAR studies reported for three known LXR ligands, epoxycholesterol[51] (CO1), GW3965[50] (965), and Tularik[50] (44 B).

In the present study, we distinguish active and inactive compounds by the drug potency (EC50) value. A threshold value of 1 μM was chosen to discriminate between active and inactive compounds in view of the distribution of all EC50 values (ranging from 0.010 to 12 μM) of the 113 compounds in the AI_DS. The choice of the threshold value came from our collaboration with experimentalists working in the field of LXRβ who, according to their own experience, proposed the value of 1 μM to discriminate the active compounds from the inactive ones. The resulting compound data sets contain 70 active and 43 inactive compounds, respectively.

Active Validation Data Set (AV_DS): Another random subset of 74 active molecules different from the ones selected in AI_DS was also extracted from the three SAR studies mentioned above in order to constitute the active validation data set (AV_DS). This data set was used in KDD studies in order to restrict the KDD-selected compounds to those which are the most similar to compounds contained in the AV_DS.

Input Collection Data Set (Input_Coll_DS): This data set is obtained by merging the Zinc_Subset_DS and AI_DS. So, the Input_Coll_DS data set contains 598 488 compounds which serve as input data in the two structure-based filtering experiments.

In order to investigate conformational spaces and to use a well-adapted independent sampling in each filtering procedure, chemical structures of Input_Coll_DS were submitted to the conformational sampling tool Omega[60] (yielding 91 million structures stored in the conformer data set called Input_Conf_DS1) and to the specific conformational sampling tool of phase[61,62] (yielding 119.6 million structures stored in the conformer data set called Input_Conf_DS2) for the shape and the pharmacophore filters, respectively.

The Filters. *Shape Filter Using MSSH/SHEF.* The so-called MSSH/SHEF shape filter used in this paper is an in-house geometrical matching procedure, namely molecular surface spherical harmonic (MSSH)/spherical harmonic coefficient filter (SHEF).The MSSH/SHEF procedure has already been described in various papers.[18,34] Briefly, each 3D structure of the Input_Conf_DS1 conformer data set was submitted to the MSSH procedure in order to describe its shape with the spherical harmonics (SH) expansion coefficients. Representation of the molecular surfaces of a target binding site and a ligand by their expansion coefficients allows a shape comparison between the two surfaces to be achieved. For this purpose, considering the surface of the target as rigid and fixed, the SHEF program will rotate the coefficients of the ligand molecule in order to obtain the minimal root-mean-square distance of these coefficients to those of the target. This was achieved within the VSM-G platform by distributing the necessary calculations onto a cluster grid.[63] Each conformer was thus associated with 10 SH coefficients representation.

The 3D Pharmacophore Filter (Pharm3D). This filter consists of two steps involving two independent programs. The 3D pharmacophore definition uses the efficient LigandScout[64,65] program to generate pharmacophore models based on the three crystallographic structures of target−ligand complexes used here (1p8d, 1pq6, and 1pq9).[40,66] In LigandScout, 3D pharmacophores are based on a defined set of seven types

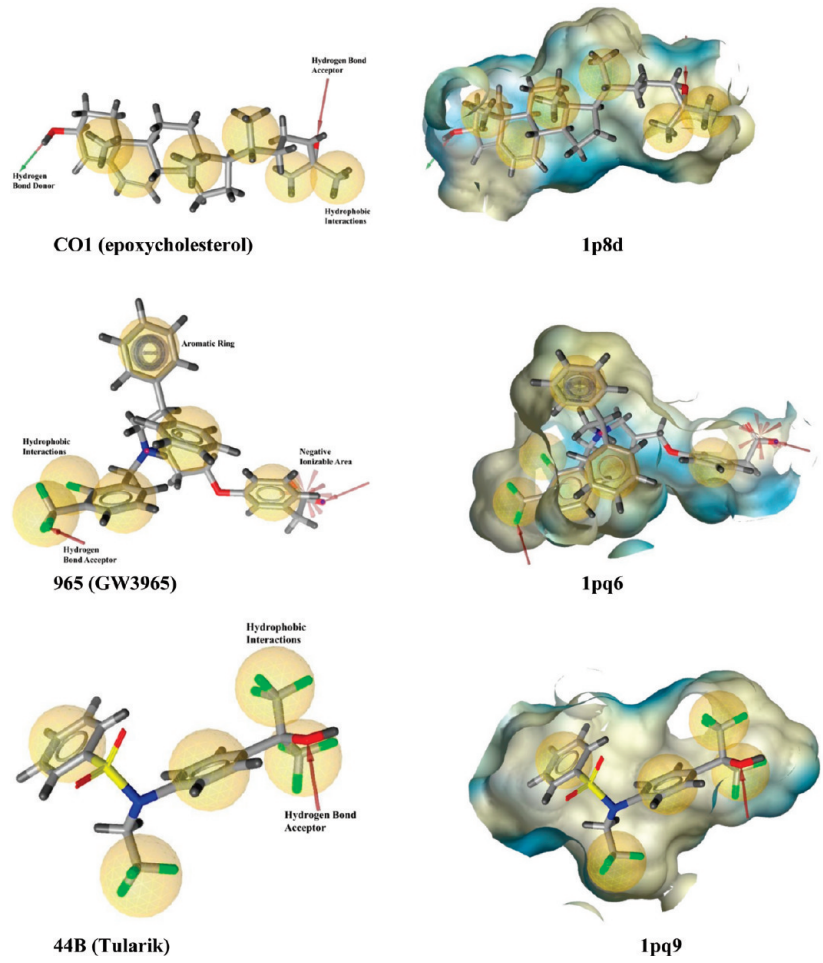

Figure 2. Pharmacophoric points of each crystallographic ligand and its binding in its cognate protein crystal structure.

of chemical features and volume constraints which are sufficiently selective to identify the binding mode. They are thus useful for establishing 3D patterns for screening databases. The pharmacophores obtained for each of the three crystallographic structures used in this study are shown in Figure 2 as well as the corresponding ligand positions within each binding site. And then, the phase program was applied in order to screen all the ligand conformers from the Input_Conf_DS2 data set, based on the three pharmacophore models produced by the LigandScout program. The phase program was selected here because it can import LigandScout data readily. For similar reason, the conformers in the Input_Conf_DS2 data set were generated with the phase program rather than the Omega program. Scoring produced by the phase program ranges from 0 to less than 1 and decreases when the matching between a conformer molecule and a pharmacophore improves.

Knowledge Discovery from Database (KDD) by Association Rules and Tanimoto Filter (AR+Tanimoto). In KDD, the initial data preparation step has an essential role. Current methods mostly yield a 2D table, where the data are represented as objects, displaying specific values for given properties. However, such single-table representations hardly reflect the complexity of the data to be considered in drug design. Consequently, it is desirable to perform data preparation in a more general way to allow a wider panel of mining methods being applied.[41] For this purpose, we first designed a relational database to integrate and represent in a nonredundant manner various descriptors of proteins, ligand, and protein−ligand interactions. We then designed a set of views capable of extracting various data sets for the data-mining step.

The KDD filter is based on the P3LI database (protein−protein and protein−ligand interactions) that we are currently developing, which is dedicated to the integration of heterogeneous data on proteins, ligands, protein−ligand, and protein−protein interactions. A key point of our approach is that this database will not only be queried for getting information but also for building data sets to be mined subsequently. Our methodology has four steps: (i) building a conceptual data model; (ii) specifying a relational data model and a workflow for collecting data on the proteins and on active and inactive ligands, requiring the design of specific wrappers; (iii) defining views on the data model for each requirement and extracting relevant patterns using data-mining methods; and (iv) using patterns retrieved as filters on a compound database and using the Tanimoto[67] score on ligand chemical fingerprint (CFp)[68] to rank the ligands with respect to their dissimilarities to a data set of active molecules.

The conceptual data model relies on the analysis of the VHTS task. Figure 3 presents the entity-relationship (ER) data model for the P3LI database, including primary data source and software names from which data elements will be available. The ER model contains six main entities, namely protein, ligand, domain, protein−ligand interaction (PLI), domain−domain interaction (DDI), and classification with subsequent relationships. A protein is described by

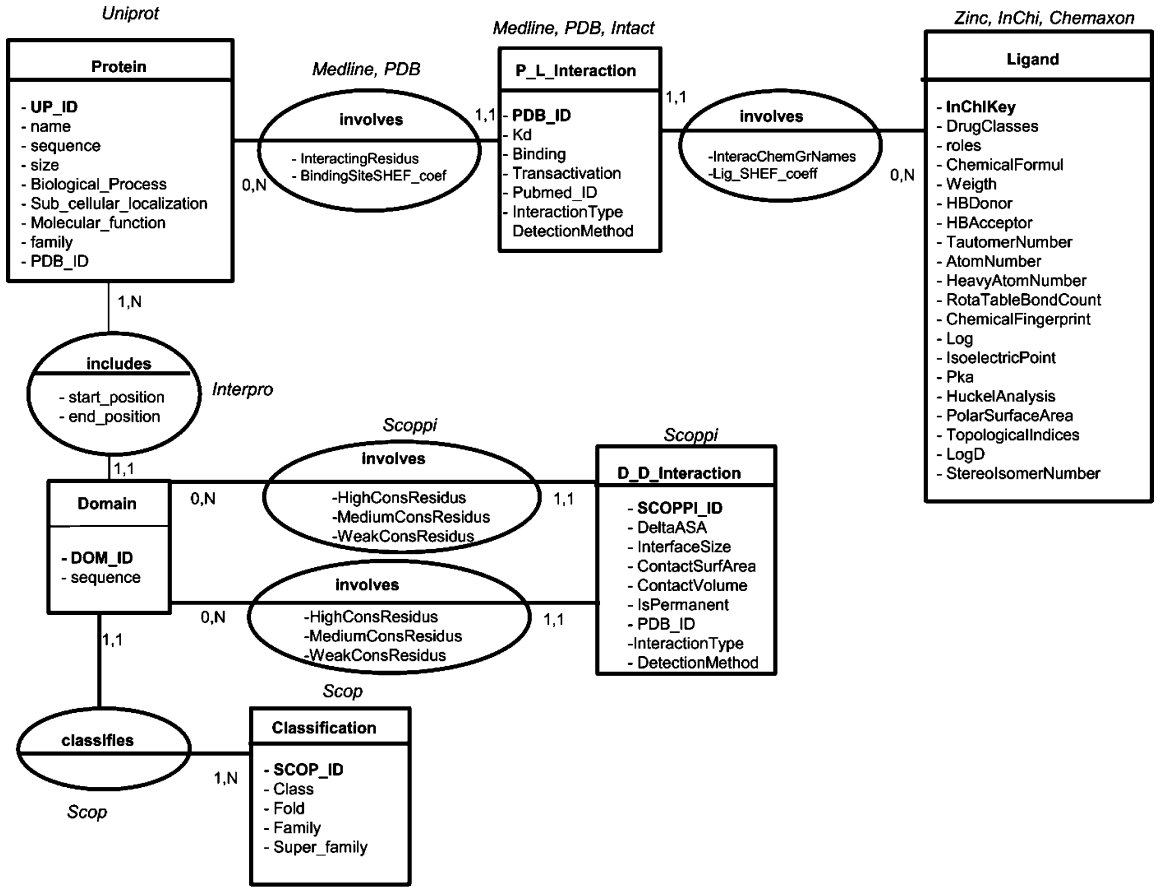

Figure 3. ER model for the P3LI[73] database, including primary data sources and software names. The ER model contains six main entities namely protein, ligand, domain, PLI, DDI, and classification with relevant relationships and attributes.

several attributes (e.g., name, sequence, and size) taken from the UniProt[69] database. A protein is composed of a set of domains (SCOP[70] classification), each of which are numbered and positioned according to the sequence of the protein (Interpro).[71] A set of physical and chemical attributes are assigned to the ligands or computed by specific software, such as ChemAxon, INCHI, or MSSH (weight, number of hydrogen-bond donors/acceptors, topological indices such as Balaban index, spherical harmonic coefficients describing the ligand surface, etc.). A protein may have identified PLI which are documented either in the PDB,[52] Pubmed, or IntAct[72] databases by a set of characteristics (e.g., K_d, binding rate, and EC50 value). Additional fields are used to describe the PLI and PPI themselves, such as the names of the interacting residues or chemical moieties.

Populating the P3LI database requires that data on a target protein be collected by successive querying of the identified data sources or by calling relevant softwares. Concerning the PLI, a list of known interacting ligands with the target protein was first built by querying PDB and PubMed with the protein name, and then most of the PLI attributes were extracted from PDB or IntAct databases.

Once the P3LI database integration is finished, various views are defined on the data model in order to produce the desired data sets for mining. We explore a set of physicochemical properties of LXR known ligands. A SQL view DS1 is defined to select the desired set of properties for active "yes" and nonactive "no" LXR ligands that are present in the P3LI database.[73]

```
CREATE VIEW DS1 (HeavyAtomCount, AtomCount, ExactMass, IsoelectricPoint,
TautomerCount, LogD, LogP, Apka, Bpka, StereoIsomerCount, RotatableBondCount,
HBAcceptorCount, HBDonorCount, PolarSurfaceArea, BalabanIndex, HararyIndex,
HyperwienerIndex,  activeLXRLigand)
 AS
 (SELECT   HeavyAtomCount, Atomcount … HyperwienerIndex, 'yes'
 FROM      Ligand li, P_L_Interaction pli, Protein p
 WHERE     p.Name = "LXR"
 AND       li.InchiKey=pli.InchiKey
 AND       p.UP_ID=pli.UP_ID
 AND       Drug potency not NULL)
 UNION
 (SELECT   HeavyAtomCount, Atomcount ,… HyperwienerIndex, 'no'
 FROM      Ligand li, P_L_Interaction pli, Protein p
 WHERE     p.Name = "LXR"
 AND       li.InchiKey=pli.InchiKey
 AND       p.UP_ID=pli.UP_ID
 AND       Drug potency not NULL)
```

A KDD filter based on the physicochemical properties of actual LXR ligands can result in regularities found in ligand properties associated to the fact of being an active LXR ligand. Such regularities can be detected by a program searching frequent item sets and association rules, such as the Apriori algorithm implemented in the Weka[74] machine learning workbench.[42] The DS1 data set used for this study corresponds to the execution of the DS1 view on the LXR P3LI database. It contains 18 properties for the 113 LXR ligands of the AI_DS data set that includes 70 active and 43 nonactive ligands according to the EC50 value. Several ligand properties display continuous numeric values and need

Table 2. Recovery of the LXR Crystallographic Ligands with Respect to the Three Known LXR Structures with Either the MSSH/SHEF Shape or the Pharm3D Filter[a]

	filter	1p8d	1pq9	1pq6
epoxycholesterol (CO1)	MSSH/SHEF	1267 (0.58)	absent	absent
	Pharm3D	1 (0.52)	absent	absent
GW3965 (965)	MSSH/SHEF	6108 (0.62)	2553 (0.61)	4 (0.46)
	Pharm3D	absent	absent	4 (0.42)
Tularik (44B)	MSSH/SHEF	absent	1036 (0.66)	absent
	Pharm3D	4670 (0.99)	2 (0.21)	absent

[a] For each cell in the table, the first number corresponds to the rank of the ligand among the top 10 K compounds selected by the corresponding filter from Input_Conf_DS1 or Input_Conf_DS2 library, and the number in brackets corresponds to its score according to the filter used. For each target, the underline numbers correspond to the ligand present in the corresponding PDB entry.

to be discretized before applying the Apriori program. It is also necessary to extrapolate several missing data by using the mean of all other values supported by this property. The Apriori program was run on the DS1 data set with a minimum confidence of 0.7. The association rules are filtered in order to retain those having a consequent (or right part) containing the property Active LXRLigand = 'yes'. Among the first 50 rules, we obtain 9 rules including the searched property. The following two rules seem to be a relevant basis for LXR ligand filtering:

(1) (StereoIsomerCount ≤ 40) AND (HyperwienerIndex ≤ 28 958) → ActiveLXRLigand ='yes' (confidence = 0.76, support = 0.79)

(2) (TautomerCount ≤ 3.8) AND (StereoIsomerCount ≤ 40) AND (HyperwienerIndex ≤ 28 958) → ActiveLXRLigand = 'yes' (confidence = 0.72, support = 0.74)

The second association rule was used in order to filter out the Input_Coll_DS data set. This rule implies the largest number of selection criteria and has the confidence value of 0.72. It applies to 74% of the ligands, as indicated by the support value. The resulting AR_subset contains 456 745 molecules (Table 1). As the molecules are not ranked in the AR_subset, the Tanimoto[67] score between ligand 2D structures was calculated between each molecule stored in the active validation data set (AV_DS) and each molecule present in the AR_subset. In order to retrieve the same number of compounds as with the other filters, various similarity threshold values were tested. Thus the compounds selected with the KDD filter eventually correspond to the AR_subset compounds that display the highest similarity to the active SAR ligands of LXR protein present in AV_DS.

Managing, Profiling, and Analyzing Results. For each filtering method, a subset of 10 000 compounds was selected and ranked for representing a data set of the "top 10 K" molecules. The number 10 000 was retained according to previous studies as containing most of the molecules of interest.[43] The compounds were stored as seven output compounds libraries called hereafter the "10 K compound DSs": Three subsets coming from the MSSH/SHEF shape filter (Output_Shape_1p8d, Output_Shape_1pq6, and Output_Shape_1pq9), three subsets coming from the Pharm3D filter (Output_Pharm_1p8d, Output_Pharm_1pq6, and Output_Pharm_1pq9), and one subset coming from the KDD/Tanimoto filter (Output_AR+Tanimoto). Each of these subsets as well as the input one (Input_Coll_DS) were studied in terms of diversity and 'drug- ' and 'lead-like' properties using the 'ScreeningAssistant' software.[75] Compounds are stored in MySQL databases, and all the operations on these databases were carried out by using SQL queries.

Hierarchical Clustering of The 10 K Compound DSs by Maximum Common Substructure. Clustering chemical structures are widely used in various stages of the drug discovery process. Traditional clustering methods are based on similarity scores. These techniques are efficient from a computational point of view, but the results obtained are often difficult to interpret. A clustering technique that results in a highly intuitive grouping of structures can be introduced by the use of the concept of maximum common substructure (MCS).[76] The MCS of two chemical graphs is the largest subgraph shared by both graphs. In the present work, this was calculated using the MCS library of the Chemaxon package.[68] The LibraryMCS method reduces the number of MCS pair computations by first estimating which two structures are likely to have large MCS. This program groups molecules according to their MCS. The result is a dendrogram where each node represents a structure common to all compounds placed below. According to the authors, this package produces interesting clusters that are different from those obtained by other chemical clustering tools.

Comparing the Efficiency of the Three Filters. Each 10 K compound DS was analyzed with respect to the enrichment in known active or inactive molecules. The enrichment of active (or inactive) ligands from AI_DS is estimated from subsets of the 10 K compound DS, which is obtained with each filtering strategy. We also expressed efficiency with the following criteria: First, a *filter factor* was defined as the ratio between the number of compound used in the input data set and the number of compounds that passed the respective filter. Second, the *efficiency* is defined as the *filter factor* divided by the amount of CPU time required.

RESULTS

Shape Filter Using MSSH/SHEF. To estimate the quality of the MSSH/SHEF shape filter, we first checked how the three known crystallographic ligand structures (epoxycholesterol, GW3965, and Tularik) were ranked in the output 10 K compound DSs. As shown in Table 2, epoxycholesterol has a score of 0.58 and is ranked at position 1267 with regard to its cognate crystal binding pocket (PDB entry: 1p8d). In other words, there are 1266 molecules that better fit to the 1p8d binding cavity according to the SHEF scoring function. The GW3965 compound has a score of 0.46 and is well ranked at position four with its respective crystal binding pocket (PDB entry: 1pq6), indicating clearly that the binding pocket in the 1pq6 conformation is optimized to the crystal ligand GW3965. However, the GW3965 ligand itself could also bind other LXR conformations, since it is ranked with

Table 3. Percentages of Common Compounds between Two Targets in the Top 10 K Compounds Selected by the MSSH/SHEF Shape Filter and Pharm3D Filter

	filter	1p8d	1pq9	1pq6
1p8d	MSSH/SHEF	100%	49.40%	15.79%
	Pharm3D	100%	1.51%	2.03%
1pq9	MSSH/SHEF	60%	100%	13.20%
	Pharm3D	1.51%	100%	2.00%
1pq6	MSSH/SHEF	15.81%	10.39%	100%
	Pharm3D	2.03%	2.00%	100%

the two other target conformations 1pq9 and 1p8d, at positions 2553 and 6108, respectively. Finally, the score of Tularik with its own crystal binding pocket (PDB entry: 1pq9) is (0.61), and this ligand is ranked at position 1036. Neither epoxycholesterol nor Tularik are found in the 10 K compound DSs of their noncognate target conformation.

Altogether these results allow us to conclude that the MSSH/SHEF shape filter effectively retains relevant compounds that match well to the target structure tested. We then investigated whether there were shared molecules among the 10 K compound DSs deriving from each LXRβ conformation. An analysis of the redundancies between the three 10 K compound DSs is presented in Table 3. The highest percentage of common compounds is found between 1p8d and 1pq9 conformations, while low percentages of common compounds are found between 1pq6 and 1p8d or 1pq9 conformations. This is consistent with the notion that the 1pq6 binding pocket seems to be specific to GW3965-like compounds or to accept compounds having markedly different scaffolds than those binding 1p8d and 1pq9.

The enrichment in active molecules (the 70 active molecules stored in the AI_DS) in the increasing subsets of each 10 K compound DS was then studied, as shown in Figure 4A. For the 1p8d target conformation, 23 active molecules (33% of the AI_DS active molecules) appear in the first 500 molecule subset of the corresponding 10 K compound DS. The enrichment rate stabilizes at around 36% in this 1p8d 10 K compound DS. However, the results are less favorable for target conformations 1pq6 and 1pq9. The 1pq6 enrichment curve shows that only 9 active molecules (13% of AI_DS active molecules) appear in the first 500 molecule subset of the corresponding 10 K compound DS. The recovery rate stabilizes at around 16% for the 1pq6 10 K compound DS. The 1pq9 enrichment curve reveals that no AI_DS active molecule appears in the first 500 molecule subset of the corresponding 10 K compound DS. The first active compound appears in the 1500 molecule subset, and

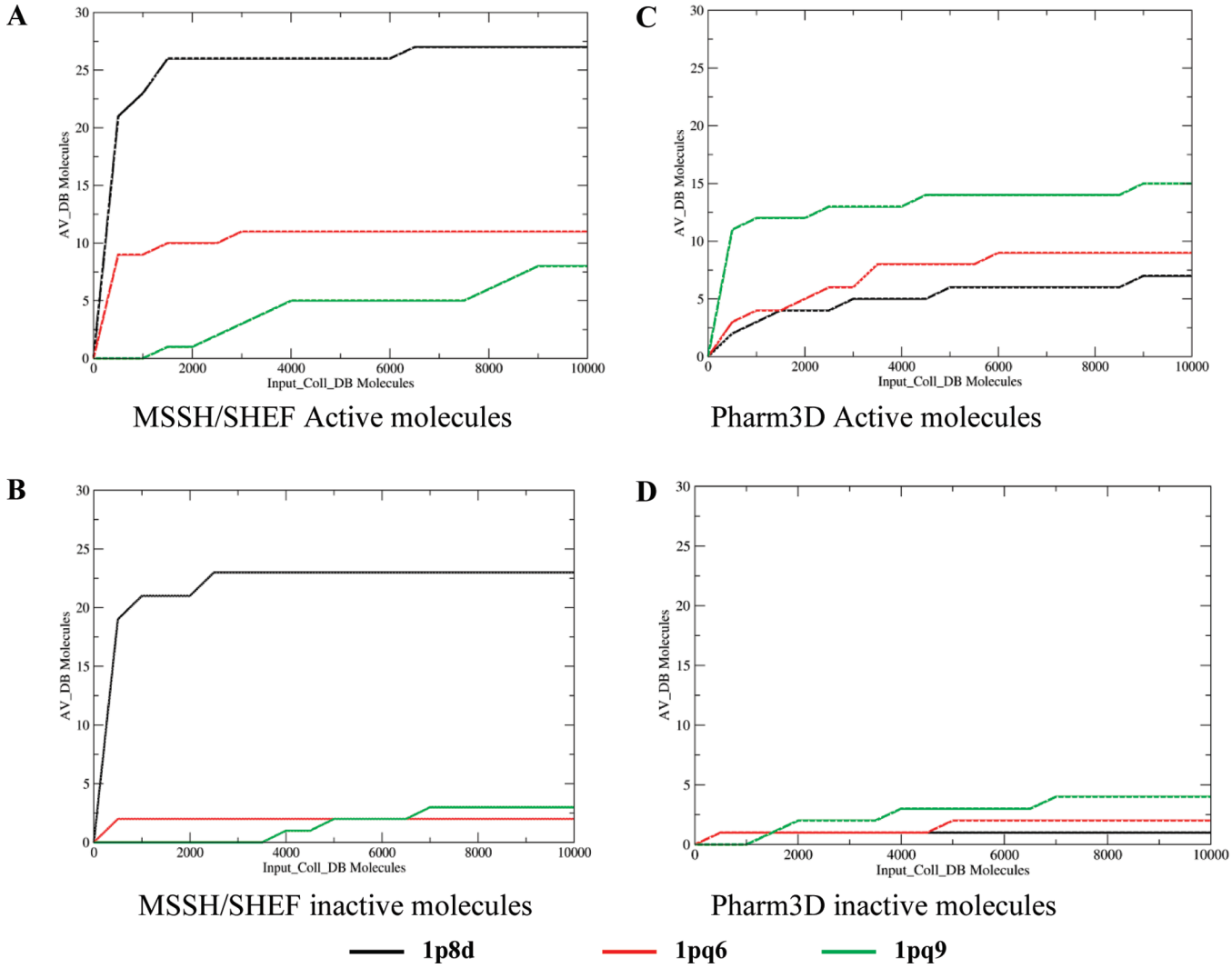

Figure 4. Enrichment in active molecules. These curves represent the number of active (A and C) and inactive (B and D) molecules from AI_DS in increasing subsets of the 10 K compounds selected by the MSSH/SHEF and Pharm3D filter.

Table 4. Analysis of the Various 10 K Compound DSs Selected by the Three Filters

10 K compound DS	number of selected compounds	number of active selected compounds	number of inactive selected compounds	number of selected crystallographic ligands	number of clustering levels	number of compounds in top-level clusters*	number of singleton clusters
Output_Pharm_1p8d	10 000	7/70	1/43	2	5	409	110
Output_Pharm_1pq6	10 000	9/70	2/43	3	5	286	40
Output_Pharm_1pq9	10 000	15/70	4/43	1	5	415	120
Output_Shape_1p8d	10 000	26/70	23/43	3	5	685	165
Output_Shape _1pq6	10 000	6/70	2/43	3	5	532	72
Output_Shape_1pq9	10 000	4/70	4/43	3	5	716	151
Output_AR+Tanimoto	10 000	69/70	41/43	3	4	83	19

the enrichment rate stabilizes at around 10% for the 1pq9 10 K compound DS.

The curves that represent the enrichment in inactive molecules (the 43 inactive molecules stored in AI_DS) are presented in Figure 4B. The 1p8d curve is very similar to the one obtained for active molecules. However, the 1pq6 and 1pq9 curves are different from those obtained for active molecules, suggesting that the MSSH/SHEF shape filter is better able to differentiate active and inactive compounds with the 1pq6 and 1pq9 LXR conformations than with the 1p8d one.

Regarding the speed and efficiency of the MSSH/SHEF shape filter (Table 5), the computing time required for filtering the 91 million conformers of Input_Conf_DS1 was 1440 min (i.e., one day). This corresponds to an efficiency of 6.3 given the filter factor of 9100.

3D Pharmacophore Filter. The results of the Pharm3D filter are summarized in Table 2. As expected, since each pharmacophore was designed from the appropriate crystallized complex structure, each crystal ligand (epoxycholesterol, GW3965, and Tularik) appears very well ranked with respect to its cognate LXR binding site. Epoxycholesterol has the best score (0.52) of all molecules tested against the 1p8d pharmacophore and, therefore, ranks first in the corresponding 10K-compound DS. However, epoxycholesterol is absent from the 10 K compound DSs obtained with the 1pq9 and 1pq6 pharmacophores. GW3965 has a score of 0.42 and ranks at position four with the 1pq6 pharmacophore, but it is absent from the 10 K compound DSs obtained with the 1p8d and 1pq9 pharmacophores. Tularik ranks second with a score of 0.21 with the 1pq9 pharmacophore compared to its score of 0.99 corresponding to a rank of 4670 with the 1p8d pharmacophore. Tularik is absent from the 10 K compound DS obtained with the 1pq6 pharmacophore. We observe here that the pharmacophore designed with the crystallographic structure 1p8d is not clearly discriminating because it can lead (with a poor score close to 1) to the selection of Tularik. The two other pharmacophores (1pq6 and 1pq9) appear by contrast very specific.

Some redundancy exists within the 10 K compound DSs selected by each different LXR structure-based pharmacophore (Table 3). Altogether the redundancy between molecules selected by different pharmacophores is no more than 2%. This suggests that each 3D pharmacophore query generates hits tightly linked to the features considered in the design of pharmacophore.

The enrichment in active and inactive molecules from AI_DS is presented in Figure 4C and D for the various 10 K compound DSs. When the 1p8d pharmacophore was used as the query, only two active molecules (2.5% of the AI_DS active molecules) were recovered within the first 500 molecule subset, with a final recovery rate at around 2.5% for the whole 10 K compound DS. The behavior of the 1pq6 pharmacophore is similar with only three active molecules found in the first 500 molecule subset (3.75%), and a stabilized recovery rate at 3.75% for the whole 10 K compound DS. A better enrichment rate is obtained with the 1pq9 pharmacophore for which 11 active molecules are recovered in the first 500 molecule subset (13.75% of the AI_DS active molecules). Nevertheless, the stabilized enrichment rate remains around 13.75% for the whole 10 K compound DS (Figure 4C).

The enrichment of the three 10 K compound DSs in AI_DS inactive molecules is very low (1−3 molecules, see Figure 4D). This implies that the pharmacophore filters can successfully differentiate between active and inactive ligands when performing a selection with the 1p8d, 1pq6, and 1pq9 pharmacophores.

Regarding the speed and efficiency of the Pharm3D filter (Table 5), the computing time required for processing the 119.6 million conformers from Input_Conf_DS2 with each Pharm3D filter was 4320 min or three days. This leads to an efficiency of 2.75 given the filter factor of 11 900.

KDD/Tanimoto Filter. After calculation of all pairwise similarities between the AR_subset and the AV_DS, a dissimilarity threshold was determined in order to select a comparable number of compounds with the AR/Tanimoto filter, as with the other filtering strategies. This threshold value was set to 0.48, and the resulting 10 K compound DS contained molecules displaying most similarity to molecules from AV_DS. This 10 K compound DS includes the three crystal ligands (epoxycholesterol, GW3965, and Tularic) of the LXR target. We, therefore, conclude that this filtering strategy retains all three crystal ligands and molecules similar to them. In the KDD/Tanimoto 10 K compound DS, 69 active molecules (from the 70 ones present in AI_DS) is recovered (Table 4). However 41 inactive molecules (from the 43 ones present in AI_DS) were also recovered. This implies that the KDD/Tanimoto filter approach cannot reliably discriminate between active and inactive compounds. This is probably due to the fact that most active and inactive compounds in AI_DS are quite similar to the three main crystal LXR ligands used in this study.

Regarding the speed and efficiency of the filter, as summarized in Table 5, this filter uses a computing time of 420 s or 7 min to process about 600 000 molecules. This corresponds to a filter factor of 60 for an efficiency of 8.57.

Table 5. Efficiency Indicators of the Considered Filters

	MSSH+SHEF shape filter	Pharm3D filter	KDD/Tanimoto filter
number of input molecules	91 000 000	119 600 000	600 000
number of selected molecules	10 000	10 000	10 000
CPU time	1440 min	4320 min	7 min
filter factor value	9100	11,900	60
efficiency value	6.3	2.75	8.57

Comparative Evaluation of Results. After the filtering process, around 70 K (70 000) molecules have been collected from the original Input_Coll_DS library. They correspond to the molecules stored in the various 10 K compound DSs obtained from the MSSH/SHEF shape filter (Output_Shape_1p8d, Output_Shape_1pq6, and Output_Shape_1pq9), the Pharm3D filter (Output_Pharm_1p8d, Output_Pharm_1pq6, and Output_Pharm_1pq9) and the KDD/Tanimoto (Output_AR+Tanimoto) subsets (see Figure 1).

In order to compare the subsets, four properties were evaluated on each 10 K compound DS.

Hierarchical Clustering of the 10 K Compound DS by Maximum Common Substructure. The LibraryMCS tool tool (LMCS)[68] of Chemaxon was used to hierarchically cluster by maximum common substructure, the chemical structures contained in the seven 10 K compound DSs resulting f from the different filters. This allowed the analysis of the different subsets obtained with each filter, and the distribution of known substructures, as found in the crystallographic data and present in AI_DS.

The number of clusters present in each of the subsets produced by the filters represents the substructure diversity of each subset. It appears from the results presented in Table 4 that the three 10 K compound DS obtained with the MSSH/SHEF shape filter present the highest chemical diversity (532−716 compounds in top-level clusters; 72−165 singleton clusters). The three 10 K compound DSs obtained with the Pharm3D filter come at the second place in terms of substructure diversity, and finally the one obtained with the KDD/Tanimoto filter displays the smallest substructure diversity. This result was predictable since the pharmacophore filter biases the selection toward molecules presenting the highest similarities with the geometrical and chemical features present in three crystal ligands. Another reason for this result is that the 70 known active molecules in AI_DS, which were used to train the KDD filter, essentially belong to the same chemical family and, therefore, display weak diversity.

Diversity. The diversity has been computed for each 10 K compound DS against the whole Input_Coll_DS data set, by using the dissimilarity step of the SCA algorithm with SSKey-3DS fingerprints of the Screening Assistant program. Two control analyses were also run: (i) the Zinc_Subset_DS as an example of rich diversity and (ii) the AI_DS data set as an example of reduced diversity. As expected, and shown in Figure 5, the Zinc_Subset_DS displays 80% diversity with respect to the whole Input_Coll_DS data set, whereas the AI_DS displays no diversity. The 10 K compound DSs selected by the MSSH/SHEF, Pharm3D, and KDD/Tanimoto filters all display an intermediate 10−15% diversity against the whole Input_Coll_DS data set.

Chemical Space. The range of chemical space covered by the 10 K compound DSs also constitutes useful information. The dissimilarity step of the SCA algorithm with SSKey-3DS fingerprints of the Screening Assistant program was used to compute the diversity for each subset. As shown in Figure 6, the control analysis performed with

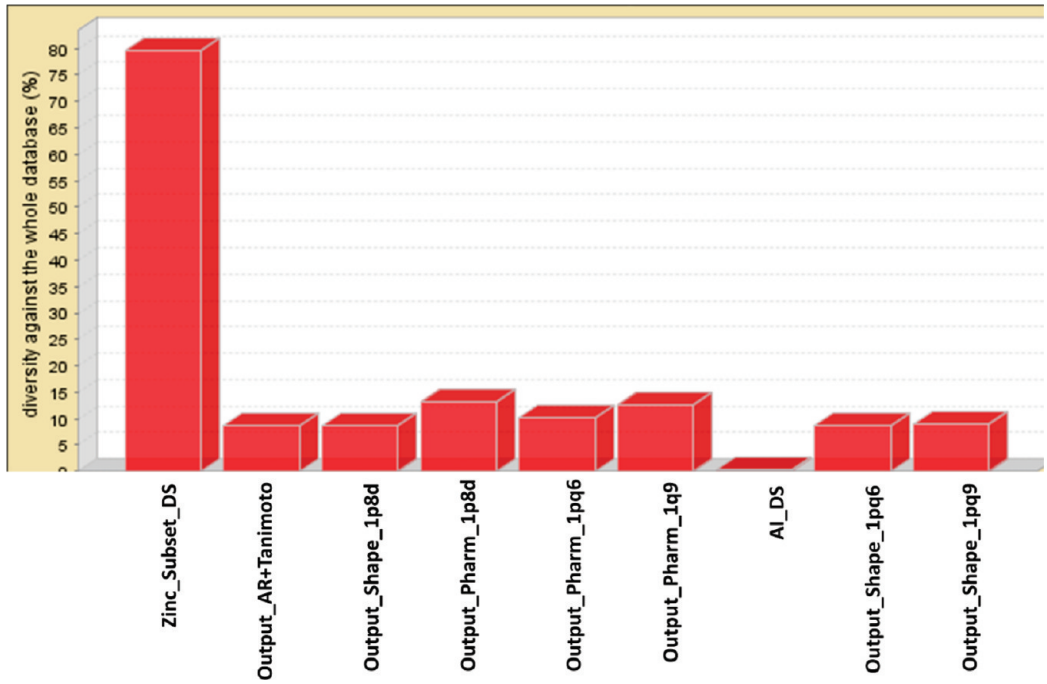

Figure 5. Global diversity. Comparison of compound diversity of the various 10 K compound DSs.

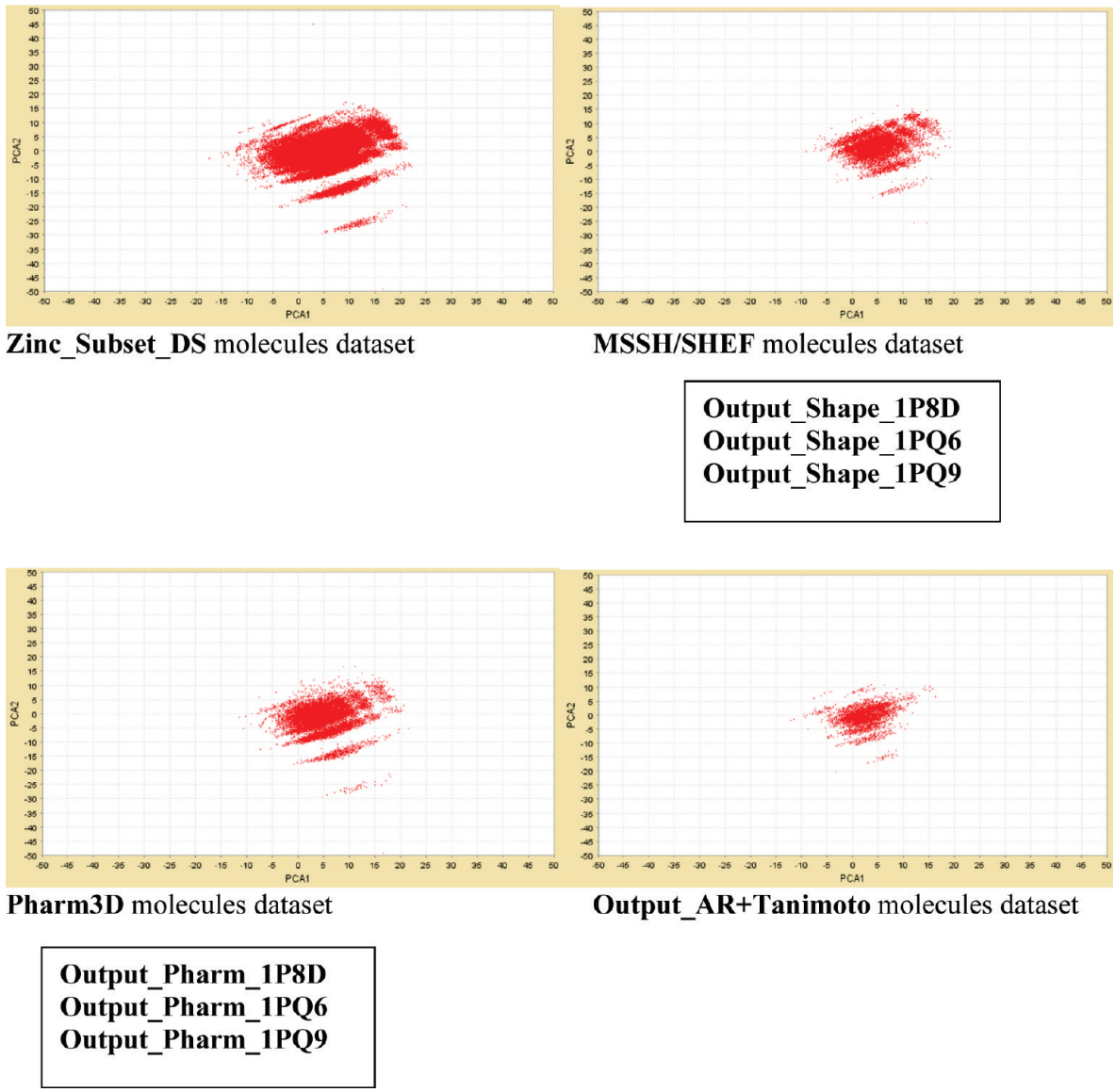

Figure 6. The range of chemical space covered by the 10 K compound DSs compared to the one covered by Zinc_subset_DS.

Zinc_Subset_DS clearly shows that this data set covers the most representative chemical space. The molecules present in each 10 K compound DS tested cover about the same chemical space as the Zinc_Subset_DS but to a smaller extent.

'Drug-Like' Properties. The 'drug-like' properties have been studied for each subset using a drug-like score provided by the Screening Assistant program. For each 10 K compound DS, the number of molecules with 0, 1, 2 and more than 2 'drug-like' failures are shown in Figure 7. All 10 K compound DSs have a high proportion of molecules with 0 or 1 'drug-like' failure. The Zinc_Subset_DS (20%), the KDD/Tanimoto 10 K compound DS (25%), and the 10 K compound DSs selected by the Pharm3D filter (from 5 to 20%) are the libraries with the highest percentages of compounds without 'drug-like' failure. It can also be noticed that the AI_DS and MSSH/SHEF contains a high percentage of molecules with two or more 'drug-like' failures. This shows that the 'drug-like' notion is not an absolute rule for filtering potential drugs.

'Lead-Like' Properties. In the early stages of a drug discovery project, it is more convenient to use 'lead' compounds. A lead compound should have a molecular weight and a log *P* smaller than a final drug compound, so that it can be optimized by adding appropriate chemical groups. Since the criteria to select 'lead-like' compounds is more stringent than those for 'drug-like', the 'lead-like' space is smaller than the 'drug-like' space. The distribution of scores obtained by each 10 K compound DS and by the two controls Zinc_Subset_DS and AI_DS using the Screening Assistant program is summarized in Figure 8. When considering the molecules with 'lead-like' failures between 0 and 1 as 'lead-like', the Zinc_Subset_DS (30%) and the compound set selected by the KDD/Tanimoto filter (40%) are the most lead-like-rich libraries. The subsets selected by the Pharm3D filters (10 to 25%) have a smaller number of 'lead-like' compounds. The compounds selected by the MSSH/SHEF shape filter, and those contained in the control AI_DS include the lowest number of 'lead like' compounds. The high percentage of 'lead-like' compounds in the Zinc_Subset_DS can be explained by the fact that this library contains many more compounds than the other subsets. By contrast, the 10 K compound DS obtained from the KDD/Tanimoto filter is

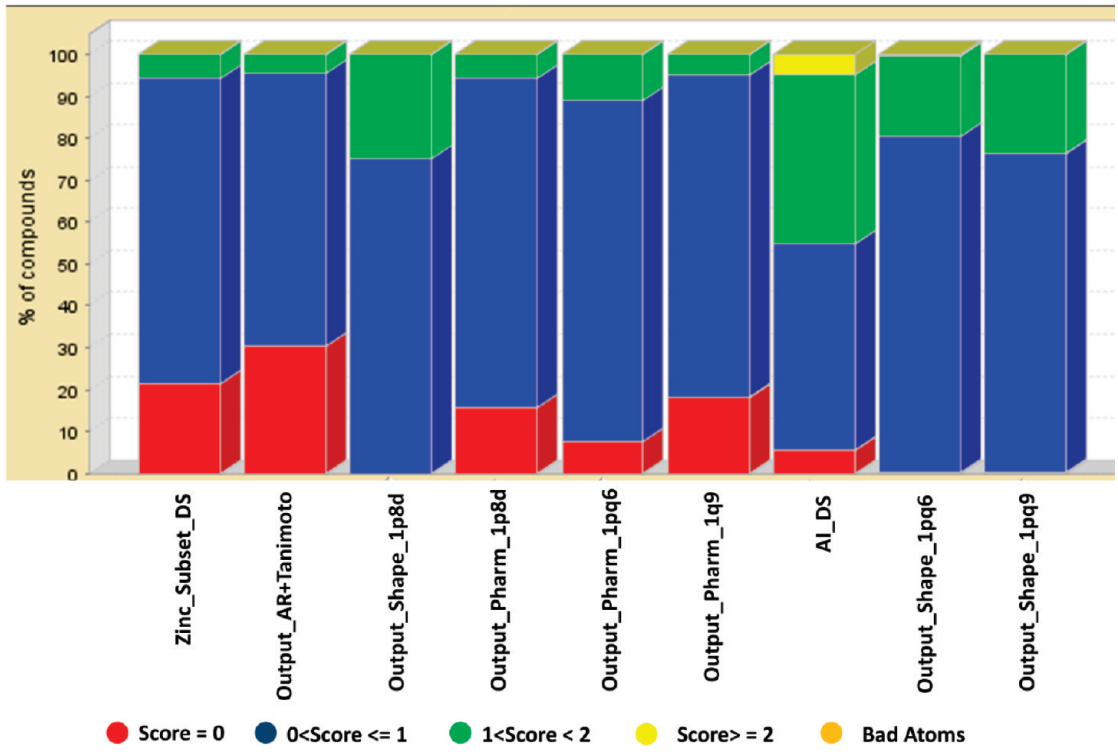

Figure 7. Drug-like score comparison of the various subsets of Input_Coll_DS.

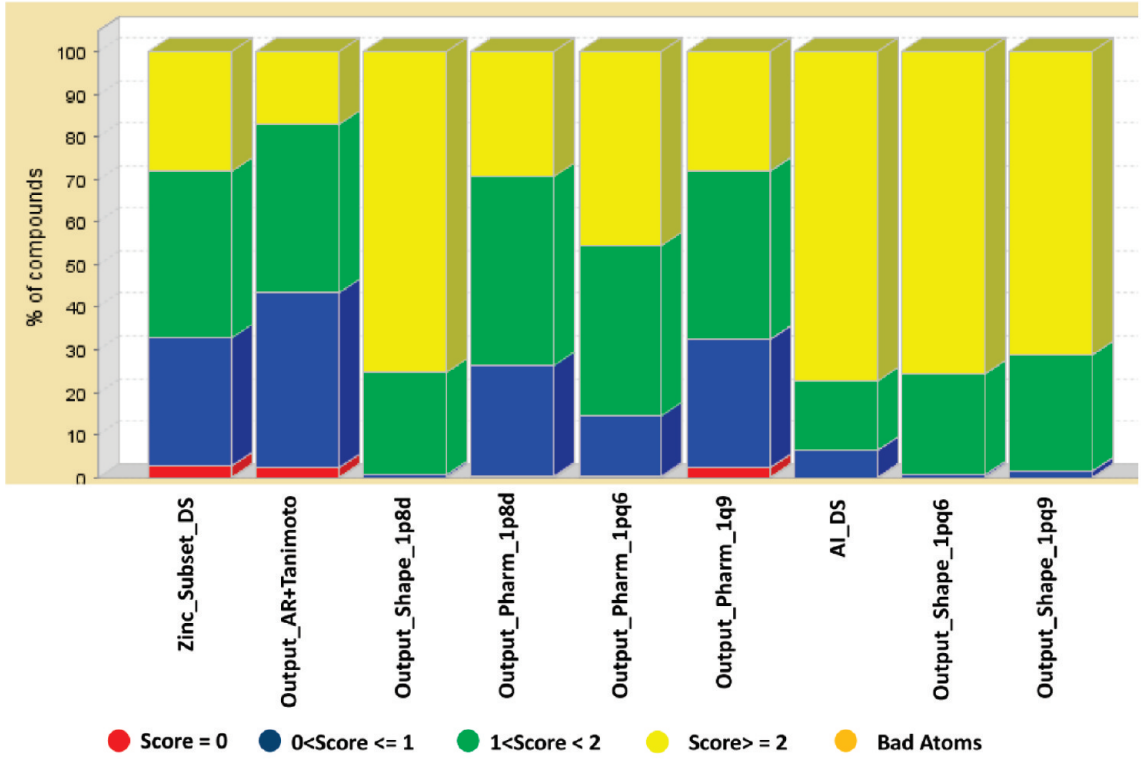

Figure 8. Lead-like score comparison of the various subsets of Input_Coll_DS.

significantly enriched in lead-like compounds compared to Zinc_Subset_DS.

DISCUSSION AND CONCLUSIONS

The work presented in this paper reveals the benefits and limitations of each investigated filter.

The MSSH/SHEF shape filter lead to the recovery of the crystal ligand of each LXRβ 3D structure tested with a good efficiency. However, each of the three crystal compounds was not ranked at the first position. The level of redundancy between the molecules recovered for each LXR conformation was about 20%. The resulting filtered libraries presented a high chemical diversity (Table 4) but a low percentage of drug- and lead-like molecules with no Lipinski rule failure (Figures 7 and 8). The spherical harmonic (SH) molecular surfaces descriptors constitute

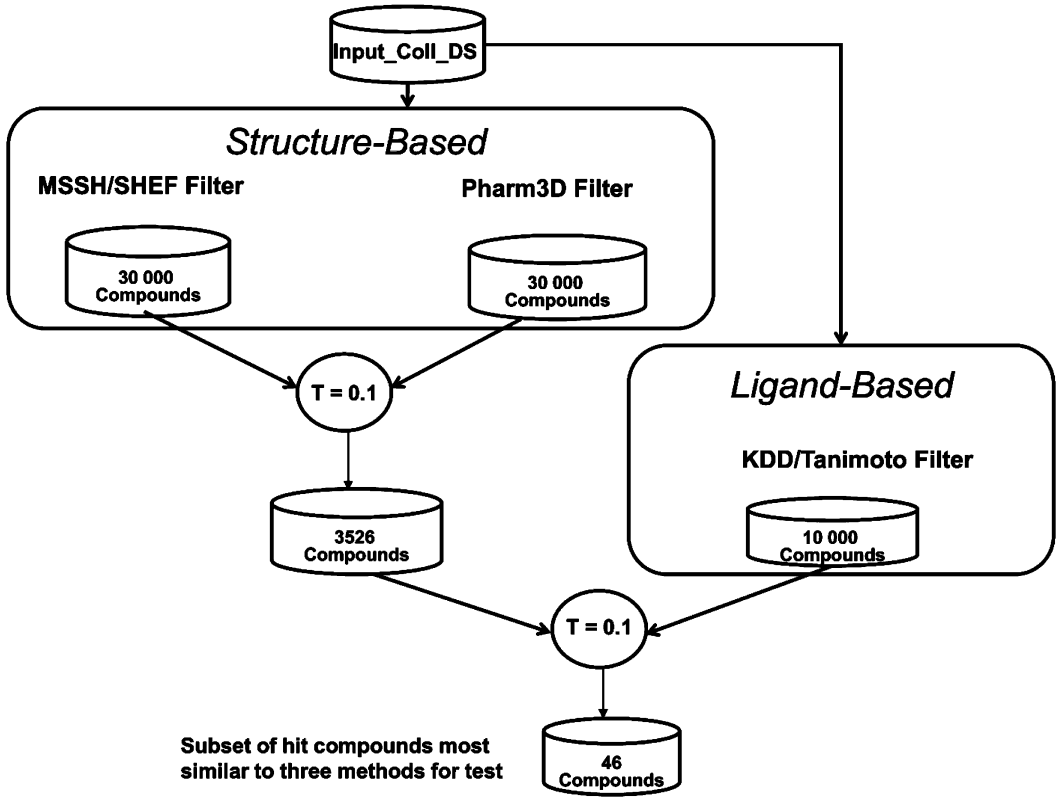

Figure 9. Proposed consensus strategy for the three filters. The identification of similar compounds between the results of the three filtering methods is based on their Tanimoto score.

a resource which can be used for other studies. Their major limitation observed in the present context lies in the inability to differentiate between known active molecules and inactive ones due to the fact that they all have very similar shapes.

The Pharm3D filter lead to the recovery of the three crystal ligands ranked at the very first positions (see Table 2). This is not surprising since the pharmacophore was established from the interactions of these compounds at their cognate binding site. A low level of redundancy (about 2%) was obtained between the molecules recovered for each of the three 1pq6, 1pq9, and 1p8d targets. Active molecules were well discriminated from the inactive ones, but a low chemical diversity was obtained (Table 4) with a high percentage of drug- and lead-like molecules with no Lipinski rule failures. The major limitation of this filter is low chemical diversity and the computation time.

The KDD/Tanimoto filter lead to the recovery among the top compounds of the three crystallographic ligands of the 1pq9, 1pq6, 1p8d targets. However, like the MSSH/SHEF shape filter, it was unable to differentiate between active and inactive molecules. The originality of this method is the reconsideration of data sets that were used in SAR studies, encompassing topological, physical, and chemical descriptors of compounds in correlation with activity data. In this study, no conformation descriptors were used with the KDD approach, neither for the ligands nor for the target. However the flexibility of the KDD approach should authorize taking into account conformation descriptors if such data can be produced or collected.[77] The KDD/Tanimoto filter produced a filtered library with a low chemical diversity (Table 4) but a high proportion of drug- and lead-like molecules with no Lipinski rule failures. This filter displayed the best efficiency due to its advantageous calculation time (Table 5). The limitation of the KDD/Tanimoto filter is inherent to its definition, since the KDD process has to be trained on a carefully selected data set representing active and inactive compounds, and this represents a strongly limiting step that can influence the resulting filtering capacity of this filter. One of the key questions appearing from the cross-comparison of filtered libraries, whatever filter was used, concerns the differences in the compounds obtained for each crystal structure of the target. Such differences can be related to the possible influence of protein flexibility during the filtering process as the three X-ray structures used in the present study are known to present significant differences in the binding site.[78] The importance of explicitly considering this flexibility during the virtual screening process is now recognized, and several ways of tackling this problem have been proposed and applied.[79–82] One common proposition is to use different crystal structures of the protein as representative samples of the conformational space.[83,84] This was partially successful, but it was shown that taking the consensus of them would improve the selection and reduce the number of false positives.[85,86]

Given that each filtering method has its own merits and drawbacks, an interesting solution could be to use all three of them in parallel and to extract common data sets from the resulting filtered libraries. This would allow combining the advantages of the different approaches and should provide a way for escaping from the various limitations encountered in several virtual screening algorithms. Hopefully this strategy would reduce the high rate of false positives, which is currently one of the main obstacles in virtual screening. Other studies led to several heuristic methods.[87–89] Most of them consider that using a consensus of several methods

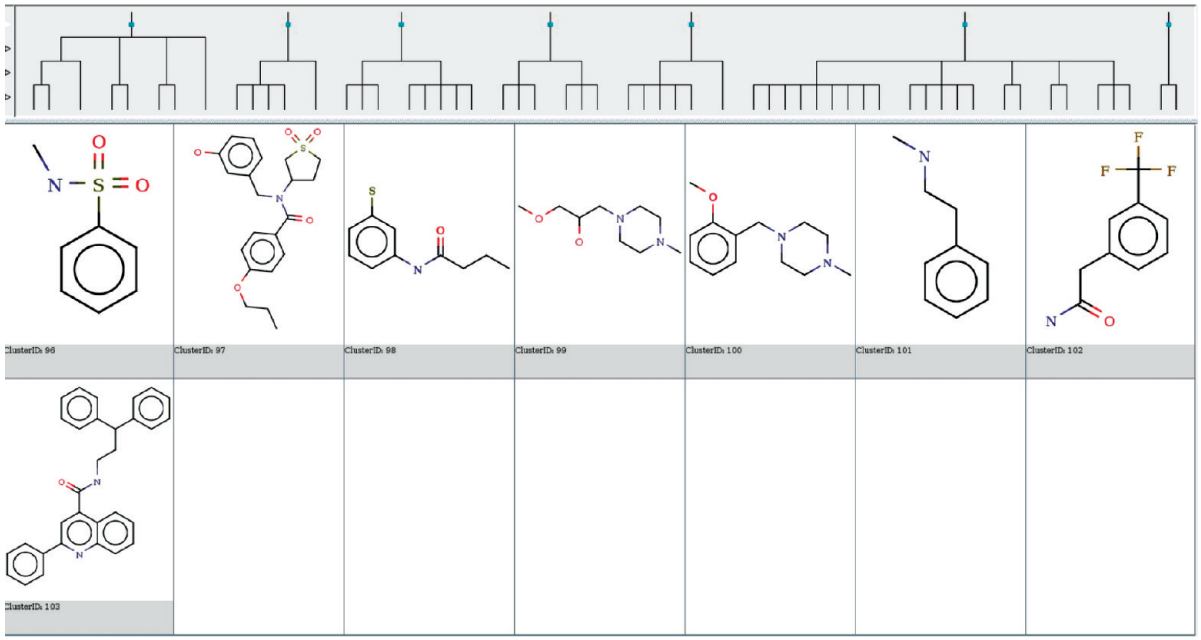

Figure 10. Illustration of the diversity of consensus compounds by substructure clustering analysis.

should improve screening efficiency. In practice, consensus approaches could have both strengths and weaknesses. One disadvantage could be the loss of active compounds rejected by a sole program. In the consensus approach proposed with the three filters presented in this study, compounds were filtered in parallel with, on the one hand two structure-based filters (MSSH/SHEF shape and Pharm3D filters) and on the other hand the KDD/Tanimoto filters. The most similar compounds of the two structured-based methods were selected according to their Tanimoto score. The resulting set was compared with compounds obtained by the KDD-based method, and the most similar compounds according to their Tanimoto score were retained. To compute the Tanimoto score, the threshold was positioned at 0.1, in order to be selective enough to keep a restricted number of compounds for the next step of drug discovery process. Thus the final filtered library was composed only of the shared compounds (Figure 9). So that, at the end, only 46 molecules were considered as having all the necessary requirements for acceptance. Interestingly, clustering analysis of this consensus subset showed a large chemical diversity (Figure 10). This consensus list is presented in the Supporting Information, and we hope it will be useful for future experimental testing.

ACKNOWLEDGMENT

The authors gratefully acknowledge D. Ritchie and P. Bladon for their advice and for correcting the manuscript. We thank Y. Asses for the useful discussions we had with her. L. Ghemtio was funded by grants from the Institut National de Recherche en Informatique et en Automatique (INRIA) and the Centre National pour la Recherche Scientifique (CNRS). V. Leroux was supported by a postdoctoral grant from the Institut National du Cancer (INCa). We thank Openeye and Chemaxon for allowing free access to their software according to an academic license and the laboratory of chemoinformatics at the Orléans University for their Screening Assistant program. This work was supported by Region Lorraine within the framework of the PRST MISN (MBI project).

Supporting Information Available: Supplement_material.xls with a list of 46 consensus molecules. This information is available free of charge via the Internet at http://pubs.acs.org/

REFERENCES AND NOTES

(1) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(2) Mestres, J. Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* **2002**, *30*, 797–799.
(3) Seifert, M. H.; Kraus, J.; Kramer, B. Virtual high-throughput screening of molecular databases. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 298–307.
(4) Bologa, C. G.; Olah, M. M.; Oprea, T. I. Chemical database preparation for compound acquisition or virtual screening. *Methods Mol. Biol.* **2006**, *316*, 375–388.
(5) Irwin, J. J. Using ZINC to acquire a virtual screening library. *Curr. Protoc. Bioinformatics* **2008**, *22*, 14.6.1−14.6.23.
(6) Cherkasov, A.; Ban, F.; Li, Y.; Fallahi, M.; Hammond, G. L. Progressive docking: a hybrid QSAR/docking approach for accelerating in silico high throughput screening. *J. Med. Chem.* **2006**, *49*, 7466–7478.
(7) Floriano, W. B.; Vaidehi, N.; Zamanakos, G.; Goddard, W. A. 3rd. HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* **2004**, *47*, 56–71.
(8) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225–242.
(9) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.
(10) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48*, 6012–6022.
(11) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235–249.
(12) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: a multistep strategy approach. *Proteins* **1999**, *36*, 1–19.
(13) Ananthan, S.; Zhang, W.; Hobrath, J. V. Recent advances in structure-based virtual screening of G-protein coupled receptors. *AAPS J.* **2009**, *11*, 178–185.
(14) Cannon, E. O.; Nigsch, F.; Mitchell, J. B. A Novel Hybrid Ultrafast Shape Descriptor Method for use in Virtual Screening. *Chem. Cent. J.* **2008**, *2*, 3.
(15) Perez-Pineiro, R.; Burgos, A.; Jones, D. C.; Andrew, L. C.; Rodriguez, H.; Suarez, M.; Fairlamb, A. H.; Wishart, D. S. Development of a novel virtual screening cascade protocol to identify potential trypanothione reductase inhibitors. *J. Med. Chem.* **2009**, *52*, 1670–1680.

(16) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinf.* **2008**, *9*, 184.
(17) Wegscheid-Gerlach, C. Chemoinformatics Approaches to Virtual Screening. Von Alexandre Varnek und Alexander Tropsha (Hrsg.). *Pharm. Unserer Zeit* **2009**, *38*, 473.
(18) Cai, W.; Xu, J.; Shao, X.; Leroux, V.; Beautrait, A.; Maigret, B. SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces. *J. Mol. Model.* **2008**, *14*, 393–401.
(19) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48*, 489–497.
(20) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward high throughput 3D virtual screening using spherical harmonic surface representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787–1796.
(21) Proschak, E.; Rupp, M.; Derksen, S.; Schneider, G. Shapelets: possibilities and limitations of shape-based virtual screening. *J. Comput. Chem.* **2008**, *29*, 108–114.
(22) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Successful shape-based virtual screening: the discovery of a potent inhibitor of the type I TGFbeta receptor kinase (TbetaRI). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4355–4359.
(23) Yamagishi, M. E.; Martins, N. F.; Neshich, G.; Cai, W.; Shao, X.; Beautrait, A.; Maigret, B. A fast surface-matching procedure for protein-ligand docking. *J. Mol. Model.* **2006**, *12*, 965–972.
(24) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: pharmacophore-based protein-ligand docking. *J. Med. Chem.* **2004**, *47*, 6804–6811.
(25) Klon, A. E.; Diller, D. J. Library fingerprints: a novel approach to the screening of virtual libraries. *J. Chem. Inf. Model.* **2007**, *47*, 1354–1365.
(26) Markt, P.; McGoohan, C.; Walker, B.; Kirchmair, J.; Feldmann, C.; De Martino, G.; Spitzer, G.; Distinto, S.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. Discovery of novel cathepsin S inhibitors by pharmacophore-based virtual high-throughput screening. *J. Chem. Inf. Model.* **2008**, *48*, 1693–1705.
(27) Mascarenhas, N. M.; Ghoshal, N. An efficient tool for identifying inhibitors based on 3D-QSAR and docking using feature-shape pharmacophore of biologically active conformation - A case study with CDK2/CyclinA. *Eur. J. Med. Chem.* **2008**, *43*, 2807–2818.
(28) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(29) Muthas, D.; Sabnis, Y. A.; Lundborg, M.; Karlen, A. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. *J. Mol. Graph. Modell.* **2008**, *26*, 1237–1251.
(30) Pandit, D.; So, S. S.; Sun, H. Enhancing specificity and sensitivity of pharmacophore-based virtual screening by incorporating chemical and shape features--a case study of HIV protease inhibitors. *J. Chem. Inf. Model.* **2006**, *46*, 1236–1244.
(31) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng., Des. Sel.* **1993**, *6*, 723–732.
(32) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 587–596.
(33) Sun, H. Pharmacophore-based virtual screening. *Curr. Med. Chem.* **2008**, *15*, 1018–1024.
(34) Cai, W.; Shao, X.; Maigret, B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graph. Modell.* **2002**, *20*, 313–328.
(35) DesJarlais, R. L.; Dixon, J. S. A shape- and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 231–242.
(36) Filikov, A. V.; Mohan, V.; Vickers, T. A.; Griffey, R. H.; Cook, P. D.; Abagyan, R. A.; James, T. L. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 593–610.
(37) Niedbala, H.; Polanski, J.; Gieleciak, R.; Musiol, R.; Tabak, D.; Podeszwa, B.; Bak, A.; Palka, A.; Mouscadet, J. F.; Gasteiger, J.; Le Bret, M. Comparative molecular surface analysis (CoMSA) for virtual combinatorial library screening of styrylquinoline HIV-1 blocking agents. *Comb. Chem. High Throughput Screening* **2006**, *9*, 753–770.
(38) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
(39) Doddareddy, M. R.; Choo, H.; Cho, Y. S.; Rhim, H.; Koh, H. Y.; Lee, J. H.; Jeong, S. W.; Pae, A. N. 3D pharmacophore based virtual screening of T-type calcium channel blockers. *Bioorg. Med. Chem.* **2007**, *15*, 1091–1105.
(40) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
(41) Hristovski, D.; Stare, J.; Peterlin, B.; Dzeroski, S. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Stud. Health Technol. Informat.* **2001**, *84*, 1344–1348.
(42) Witten, I. H.; Frank, E. Output:knowledge representation. In *Data Mining: Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005; Vol. 11, pp 61−82.
(43) Beautrait, A.; Leroux, V.; Chavent, M.; Ghemtio, L.; Devignes, M. D.; Smail-Tabbone, M.; Cai, W.; Shao, X.; Moreau, G.; Bladon, P.; Yao, J.; Maigret, B. Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment. *J. Mol. Model.* **2008**, *14*, 135–148.
(44) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discovery Today* **2000**, *5*, 61–69.
(45) Toba, S.; Srinivasan, J.; Maynard, A. J.; Sutter, J. Using pharmacophore models to gain insight into structural binding and virtual screening: an application study with CDK2 and human DHFR. *J. Chem. Inf. Model* **2006**, *46*, 728–735.
(46) Simmons, K.; Kinney, J.; Owens, A.; Kleier, D. A.; Bloch, K.; Argentar, D.; Walsh, A.; Vaidyanathan, G. Practical outcomes of applying ensemble machine learning classifiers to High-Throughput Screening (HTS) data analysis and screening. *J. Chem. Inf. Model.* **2008**, *48*, 2196–2206.
(47) Lala, D. S. The liver X receptors. *Curr. Opin. Investig. Drugs* **2005**, *6*, 934–943.
(48) Morello, F.; de Boer, R. A.; Steffensen, K. R.; Gnecchi, M.; Chisholm, J. W.; Boomsma, F.; Anderson, L. M.; Lawn, R. M.; Gustafsson, J. A.; Lopez-Ilasaca, M.; Pratt, R. E.; Dzau, V. J. Liver X receptors alpha and beta regulate renin expression in vivo. *J. Clin. Invest.* **2005**, *115*, 1913–1922.
(49) Collins, J. L. Therapeutic opportunities for liver X receptor modulators. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 692–702.
(50) Farnegardh, M.; Bonn, T.; Sun, S.; Ljunggren, J.; Ahola, H.; Wilhelmsson, A.; Gustafsson, J. A.; Carlquist, M. The three-dimensional structure of the liver X receptor beta reveals a flexible ligand-binding pocket that can accommodate fundamentally different ligands. *J. Biol. Chem.* **2003**, *278*, 38821–38828.
(51) Williams, S.; Bledsoe, R. K.; Collins, J. L.; Boggs, S.; Lambert, M. H.; Miller, A. B.; Moore, J.; McKee, D. D.; Moore, L.; Nichols, J.; Parks, D.; Watson, M.; Wisely, B.; Willson, T. M. X-ray crystal structure of the liver X receptor beta ligand binding domain: regulation by a histidine-tryptophan switch. *J. Biol. Chem.* **2003**, *278*, 27138–27143.
(52) 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.
(53) Bennett, D. J.; Carswell, E. L.; Cooke, A. J.; Edwards, A. S.; Nimz, O. Design, structure activity relationships and X-Ray co-crystallography of non-steroidal LXR agonists. *Curr. Med. Chem.* **2008**, *15*, 195–209.
(54) Janowski, B. A.; Grogan, M. J.; Jones, S. A.; Wisely, G. B.; Kliewer, S. A.; Corey, E. J.; Mangelsdorf, D. J. Structural requirements of ligands for the oxysterol liver X receptors LXRalpha and LXRbeta. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 266–271.
(55) Spencer, T. A.; Li, D.; Russel, J. S.; Collins, J. L.; Bledsoe, R. K.; Consler, T. G.; Moore, L. B.; Galardi, C. M.; McKee, D. D.; Moore, J. T.; Watson, M. A.; Parks, D. J.; Lambert, M. H.; Willson, T. M. Pharmacophore analysis of the nuclear oxysterol receptor LXRalpha. *J. Med. Chem.* **2001**, *44*, 886–897.
(56) *Chemdiv*, The chemistry of cures; Chemdiv, Inc.: San Diego, CA; http://chemdiv.emolecules.com. Accessed November 30, 2009.
(57) *Enamine*, Smart chemistry solutions; Enamine Ltd.: Kiev, Ukraine; http://www.enamine.net. Accessed November 30, 2009.
(58) AMRI, Chemical compound database; Albany Molecular Research, Inc.: Albany, NY; http://www.amridirect.com. Accessed November 30, 2009.
(59) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
(60) *OpenEye*, version 2.2.1; OpenEye Scientific Software: Santa Fe, NM, 2009.
(61) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(62) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching. *Chem. Biol. Drug Des.* **2006**, *67*, 370–372.
(63) Raphaël Bolze, F. C.; Eddy, Caron; Michel Daydé, Frederic Desprez; Emmanuel, Jeannot; Yvon, Jégou; Stéphane, Lanteri; Julien, Leduc; Noredine, Melab; Guillaume, Mornet; Raymond, Namyst; Pascale, Primet; Benjamin, Quetier; Olivier, Richard; El-Ghazali, Talbi; Touché, Irena. Grid'5000: a large scale and highly reconfigurable experimental grid testbed. *Int. J. High Perform. Comput. Appl.* **2006**, *20*, 481–494.
(64) Wolber, G.; Dornhofer, A.; Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 773–788.
(65) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2004**, *45*, 160–169.
(66) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13*, 23–29.
(67) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163–166.
(68) *JKlustor*, version 5.2.6; Chemaxon: Budapest, Hungary, 2009.
(69) Consortium, T. U. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **2008**, *37*, 169–174.
(70) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **2008**, *36*, D419−D425.
(71) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Buillard, V.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Daugherty, L.; Dibley, M.; Finn, R.; Fleischmann, W.; Gough, J.; Haft, D.; Hulo, N.; Hunter, S.; Kahn, D.; Kanapin, A.; Kejariwal, A.; Labarga, A.; Langendijk-Genevaux, P. S.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Orengo, C.; Petryszak, R.; Selengut, J. D.; Sigrist, C. J.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H.; Yeats, C. New developments in the InterPro database. *Nucleic Acids Res.* **2007**, *35*, D224−D228.
(72) Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; Kohler, C.; Khadake, J.; Leroy, C.; Liban, A.; Lieftink, C.; Montecchi-Palazzi, L.; Orchard, S.; Risse, J.; Robbe, K.; Roechert, B.; Thorneycroft, D.; Zhang, Y.; Apweiler, R.; Hermjakob, H. IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.* **2007**, *35*, D561–D565.
(73) Ghemtio, L.; Bresso, E.; Souchet, M.; Maigret, B.; Smaïl-Tabbone, M.; Devignes, M.-D. Model-driven data integration for mining protein-ligand and protein-protein interactions in a drug design context. In *Proceedings of the 9th Open Days in Biology, Computer Science and Mathematics*; Journées Ouvertes Biologie Informatique Mathématiques, Lille, France, June 30−July 2, 2008; INRIA: Lille, France, 2008.
(74) Mark Hall, E. F.; Geoffrey, Holmes; Bernhard, Pfahringer; Peter, Reutemann; Ian H., Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11*, 10–18.
(75) Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Diversity* **2006**, *10*, 389–403.
(76) Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Rault, S. The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1043–1052.
(77) Ghemtio, L.; Smail-Tabbone, M.; Devignes, M.-D.; Souchet, M.; Maigret, B.; et al. A KDD Approach for Designing Filtering Strategies to Improve Virtual Screening. In *KDIR* - International Conference on Knowledge Discovery and Information Retrieval, Madeira, Portugal, October 5−8, 2009; Ana, F., Ed., INSTIC: Madeira, 2009.
(78) Beautrait, A.; Karaboga, A. S.; Souchet, M.; Maigret, B. Induced fit in liver X receptor beta: a molecular dynamics-based investigation. *Proteins* **2008**, *72*, 873–882.
(79) Bolstad, E. S.; Anderson, A. C. In pursuit of virtual lead optimization: pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins* **2009**, *75*, 62–74.
(80) C, B. R.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discovery Today* **2009**, *14*, 394–400.
(81) Davis, I. W.; Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* **2009**, *385*, 381–392.
(82) Fischer, B.; Merlitz, H.; Wenzel, W. Receptor flexibility for large-scale in silico ligand screens: chances and challenges. *Methods Mol. Biol.* **2008**, *443*, 353–364.
(83) Huang, S. Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* **2007**, *66*, 399–421.
(84) Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.
(85) Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. Molecular Docking Screens Using Comparative Models of Proteins. *J. Chem. Inf. Model.* **2009**, *49*, 2512–2527.
(86) Fukunishi, Y.; Mikami, Y.; Kubota, S.; Nakamura, H. Multiple target screening method for robust and accurate in silico ligand screening. *J. Mol. Graph. Modell.* **2006**, *25*, 61–70.
(87) Perola, E. Minimizing false positives in kinase virtual screens. *Proteins* **2006**, *64*, 422–435.
(88) Wolf, A.; Zimmermann, M.; Hofmann-Apitius, M. Alternative to consensus scoring--a new approach toward the qualitative combination of docking algorithms. *J. Chem. Inf. Model.* **2007**, *47*, 1036–1044.
(89) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

CI900356M