Journal of
CHEMICAL INFORMATION
AND MODELING

Perspective

pubs.acs.org/jcim

Recognizing Pitfalls in Virtual Screening: A Critical Review

Thomas Scior,*[,†] Andreas Bender,[‡] Gary Tresadern,[§] José L. Medina-Franco,[⊥]
Karina Martínez-Mayorga,[⊥] Thierry Langer,[∥] Karina Cuanalo-Contreras,[†] and Dimitris K. Agrafiotis[○]

[†]Pharmacy Department, Facultad de Ciencias Químicas, Universidad Autónoma de Puebla, Puebla, Pue, México

[‡]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

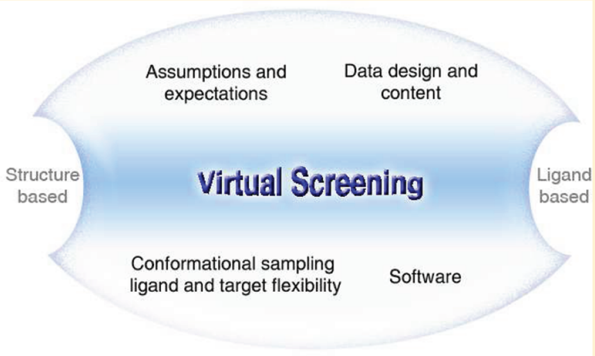
[§]Research Informatics & Integrative Genomics, Janssen Research & Development, Calle Jarama 75, Poligono Industrial, Toledo 45007, Spain

[⊥]Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, Florida 34987, United States

[∥]Prestwick Chemical, Blvd Gonthier d'Andernach, F-67400 Illkirch, France

[○]Johnson & Johnson Pharmaceutical Research & Development, LLC, Welsh & McKean Roads, Spring House, Pennsylvania 19477, United States

ABSTRACT: The aim of virtual screening (VS) is to identify bioactive compounds through computational means, by employing knowledge about the protein target (structure-based VS) or known bioactive ligands (ligand-based VS). In VS, a large number of molecules are ranked according to their likelihood to be bioactive compounds, with the aim to enrich the top fraction of the resulting list (which can be tested in bioassays afterward). At its core, VS attempts to improve the odds of identifying bioactive molecules by maximizing the true positive rate, that is, by ranking the truly active molecules as high as possible (and, correspondingly, the truly inactive ones as low as possible). In choosing the right approach, the researcher is faced with many questions: where does the optimal balance between efficiency and accuracy lie when evaluating a particular algorithm; do some methods perform better than others and in what particular situations; and what do retrospective results tell us about the prospective utility of a particular method? Given the multitude of settings, parameters, and data sets the practitioner can choose from, there are many pitfalls that lurk along the way which might render VS less efficient or downright useless. This review attempts to catalogue published and unpublished problems, shortcomings, failures, and technical traps of VS methods with the aim to avoid pitfalls by making the user aware of them in the first place.


Assumptions and expectations
Data design and content
Structure based
Virtual Screening
Ligand based
Conformational sampling ligand and target flexibility
Software

INTRODUCTION

Virtual screening (VS) is an increasingly used method to guide the identification of novel hits from large chemical libraries. VS can be divided into two broad categories, namely ligand-based and structure-based.[1] Ligand-based approaches utilize structure–activity data from a set of known actives in order to identify candidate compounds for experimental evaluation.[2] Ligand-based methods include approaches such as similarity and substructure searching, quantitative structure–activity relationships (QSAR), and pharmacophore- and three-dimensional shape matching.[3] Structure-based VS, on the other hand, utilizes the three-dimensional (3D) structure of the biological target (determined either experimentally through X-ray crystallography or NMR, or computationally through homology modeling) to dock the candidate molecules and rank them based on their predicted binding affinity or complementarity to the binding site. VS methodologies and success stories have been extensively documented in a number of recent reviews,[4−7] and the status of the field has been critically assessed.[8] The present work is focused solely on the *pitfalls of VS* and how they relate to many of the methods discussed in these references. Basic concepts of ligand-based VS have found their way into textbooks, usually presented in a nontechnical manner in order to reach a broader medicinal chemistry readership.[9] For details about VS algorithms, one recent work[9] is highly recommended.

VS can be seen as the mining of chemical spaces with the aim to distinguish molecules that possess a desired property from those that do not. As with most predictive methods, it should not be assumed that VS achieves this goal in a bullet-proof and error-free way. VS is highly dependent on the quantity and quality of available data (an aspect that was reviewed recently[10]) and the predictive/discriminatory ability of the underlying algorithm. The latter becomes even more relevant today given the plethora of available VS tools and methodologies.

Scientists working in the field of medicinal chemistry, particularly computational chemists, must not underestimate

Received: November 4, 2011
Published: March 21, 2012

the difficulties and inherent limitations of VS. QSAR offers an instructive analogy. Many of the pitfalls of QSAR, which have been published earlier,[11,12] also apply to VS (see Table 1). A fundamental assumption inherent in QSAR and pharmacophore-based VS is the "similar property principle", that is, the general observation that molecules with similar structure are likely to have similar properties. While this assumption holds true in many cases, there are many counter-examples in the field of QSAR which lead to erroneous predictions and can shake the confidence of the experimental community in the prospective utility of QSAR modeling. Interestingly, this has not yet (or not to the same extent) been the case with VS. The difference is that QSAR is typically employed to evaluate a limited number of synthetic candidates, where errors are more noticeable and costly. However, when these techniques are applied on a massive scale to screen large chemical libraries, errors are much more easily tolerated as the objective is to increase the number and diversity of hits over what would have been otherwise a random selection (see also Table 1).

In the remainder of this article, we review many of the known and still unreported limitations and technical traps of VS techniques as they are employed today. The cases are grouped together and presented as pitfalls along with possible solutions, if they exist. To assist the reader, the information is also provided in tabular format. Table 2 summarizes VS drawbacks that were previously reported in VS benchmark studies, with specific topics listed in Table 3.

We classified pitfalls into four categories: (1) those concerning erroneous assumptions and expectations; (2) those concerning data design and content; (3) those relating to the choice of software; and (4) those concerning conformational sampling as well as ligand and target flexibility.

1. PITFALLS CONCERNING ERRONEOUS ASSUMPTIONS AND EXPECTATIONS

a. Pitfall: Expectation of Identifying High-Affinity Compounds by VS. The main goal of VS is to identify novel bioactive chemical matter for the particular target of interest.[8] However, sometimes experimentalists and peer-reviewers expect VS to identify highly potent compounds. Although high potency hits are, of course, *desirable*, this should not be a *condition* to consider a VS protocol successful since affinity is typically optimized in the hit-to-lead and lead optimization stages. In this regard, VS should not be viewed differently than HTS; the main goal is to yield as many and as diverse starting points as possible, not as potent as possible. Setting a binary activity threshold for assessing VS algorithms is difficult and arbitrary. This pitfall of evaluating VS methods relates both to the choice of benchmark data in comparative studies (where some of the data sets used, such as those from the MDL Drug Data Repository, often do not contain numerical activity values) as well as to the prospective utility in pharmaceutical screening (where results can be very different depending on the activity threshold chosen).

b. Pitfall: Stringency of Queries. When employing pharmacophore models, a fundamental question that needs be tackled is the uncertainty allowed in the relative positions of the pharmacophoric features in 3D space. On the one hand, very strict settings would lead to poor structural diversity in the compounds retrieved from VS, while on the other a very fuzzy model is more likely to return a large number of false positives. Again, the settings chosen are largely arbitrary and may depend on the experience of the user performing the search.

Table 1. QSAR-Related Pitfalls in VS Described in Detail in a Previously Published Review[11] [a]

issue or keywords	comments
One at a time approach	The assumption of additive contributions to biological activity from the chemical substituents of different molecules dominates medicinal chemistry and drug design approaches. Modeling in the VS domain, as well as QSAR, may need to account for the potential breakdown of this behavior.
Multiple binding modes	It is usual to assume that multiple binding modes do not occur and 3D searches may find only structures with a similar binding mode.
Inactive prodrug-based VS	The accidental use of prodrugs in QSAR or VS can lead to the identification of analogues of a molecule not responsible in its entirety for the observed biological activity.
Experimental errors and inappropriate bioassays	Literature data sets are often built from different sources, where different assay procedures and detection techniques may be used. Such data sets are often used in VS but usually in a discriminant form, to flag active or inactive molecules, where the impact of inaccuracies may be less than in QSAR. This is discussed in more detail in the pitfalls concerning data design and content.
Data size and variety and test set composition	Skewed data sets, which lack sufficient chemical diversity, or actives which are too easily identifiable among the inactives can all impact VS, in particular retrospective experiments. Refer to the section described herein entitled Pitfalls Concerning Data Design and Content.
Incorrect feature selection for pharmacophore design	Incorrect feature definition can be detrimental to the outcome of VS. This is discussed in the choice of software section.
Robust VS procedures and use of software as black box	Inexperience with certain software can lead to poor application and lack of understanding and interpretation of results. In essence, treating software as a black box is typically detrimental in all aspects of molecular modeling, QSAR or VS.
Starting geometries in 3D VS	Aspects relating to choice of 3D conformation to use as a query in VS apply to 3D QSAR as well as VS. See the section herein regarding the pitfalls of conformational sampling and ligand and target flexibility.
Biased validation by unfair choice of control compounds	This also relates to inappropriate choice of a data set used to build or validate a model. Refer to two sections pitfalls concerning data design and content and choice of actives and inactives.
Multiple solutions	Multiple models in the case of QSAR or software approaches and descriptors in the case of VS exist. Each captures different characteristics of molecular similarity. It is often difficult to identify a preferred method, and so, it is often necessary to account for several. This is discussed in the section herein entitled single predictors versus ensembles.

[a] Freely available from http://www.benthamdirect.org/.

Table 2. Listing of Recently Published Work Concerning Pitfalls and Benchmarking for VS

pub year	title	key aspects and refs
2002[a]	Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools	Evaluation of tools (Catalyst, Confort, Flo99, MacroModel, and Omega) for conformational search to reproduce the receptor-bound conformations of 32 small, flexible, and drug-like ligands of crystal complexes. Results are considered identical to the crystal conformation if the rmsd is less than 0.25−0.30 Å. The best algorithm found was the Low-Mode conformational search of MacroModel.[13]
2002[a]	Can we separate active from inactive conformations?	Evaluation of 3D descriptors (filters) to separate active from inactive conformations. To this end, conformations were computed and compared with the active one extracted from a PDB entry. The best-performing descriptors are: solvent accessible surface area, number of internal interactions, and radius of gyration.[14]
2004	Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening	Evaluation of unattended and user design of pharmacophore models for VS of influenza virus neuraminidase ligands built by Catalyst. They were validated against crystal complexes with a benchmark set of positive and negative controls extract from literature. As a result, the automatically generated models could not cope with multiple function interactions on the very same atoms. In addition, they showed lower filtering power.[15]
2004	Virtual screening of chemical libraries	A revision of some implications concerning VS. For instance, the difficulty to calculate ligand−receptor binding energies, the inherent complexity of receptor structures, and the spatial limitations of conformational calculations. Nevertheless, VS is successfully applied also in comparison to HTS.[16]
2004	Docking versus pharmacophore model generation: a comparison of high-throughput virtual screening strategies for the search of human rhinovirus coat protein inhibitors	Evaluation of two VS methods based on either docking or pharmacophore features and applied for the search of antiviral candidates for human rhinovirus coat protein. Docking was performed with LigandFit and the pharmacophore screening conducted by Catalyst. Both methods were calibrated against the same benchmark set. Docking is in general more time-consuming than pharmacophore based screening but interpretation of results is easier in the latter case. It outperformed docking as to selectivity conformational space coverage.[17]
2004	Virtual combinatorial chemistry and in silico screening: efficient tools for lead structure discovery?	Review of some common VS methods. In a fully unattended way *LigandScout* extracts information about the binding mode of ligand−receptor complexes and generates pharmacophore models. It can be combined with *iLib* diverse which is a tool to create focused virtual compound libraries.[18]
2005[a]	Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms	Assessment for the best conformational search protocol using Catalyst's conformational search procedure with 510 ligand−protein complexes. The best protocol showed an rmsd of 0.93 between the computed conformation and the observed bioactive conformation. The performance is largely dependent on the number of rotatable bonds and ligand size. The potential energies of the bioactive conformers lie high above the global minima.[19]
2005	A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication	Evaluation of the use of simple molecular descriptors (e.g., number of atoms per molecule) in VS. It could retrieve known active compounds with novel scaffolds. Hence, atom counts appear to be a surprisingly effective descriptor for VS which could serve as a "negative control" (baseline performance) for the validation of novel algorithms.[20]
2005	Considerations in compound database preparation−"hidden" impact on virtual screening results	This work highlights how different treatment of chemical structures at the time of database preparation can have a significant influence on different VS scenarios. Parameters discussed include SMILES representation, stereochemical information, protonation state, and conformational searching.[21]
2005	LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters	Construction of structure-based pharmacophore models using six chemical features and presentation of LigandScout for automatic elucidation of pharmacophore information. Then the pharmacophore models successfully underwent selectivity tests.[22]
2006[a]	Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations	Evaluation of tools (Catalyst and Omega) for conformational search to reproduce the receptor-bound conformations of 778 drug ligands from crystal complexes. The rmsd between generated and bioactive conformers was the metric for benchmarking. Both tools were even in conformational generation. Omega was faster but Catalyst found more bioactive conformations.[23]
2006	Similarity-based virtual screening using 2D fingerprints	Review of 2D similarity metrics for VS. Molecules that are structurally similar are likely to have similar properties. The use of 2D fingerprints to perform VS is discussed. The fusion of similarity coefficients can improve the screening results. It was found that the Tanimoto coefficient is the best metric for computing molecular similarities.[24]
2006	Docking and scoring—theoretically easy, practically impossible?	Critical revision of docking and scoring functions for virtual HTS. VS depends on correct input preparations, like the virtual library when converting 2D drawings into 3D structures; or receptors with their protonation states or ligand tautomers. Critical for docking is also receptor flexibility and biologically relevant water molecules. The scoring functions should discriminate between correct and incorrect binding poses of a ligand, as well as active and inactive molecules. However, scoring and docking software still have limitations and therefore must be improved.[12]
2007[a]	Conformational sampling of bioactive molecules: a comparative study	Evaluation of the sampling ability for conformational space of Catalyst, Macromodel, Omega, MOE, Rubicon, and stochastic proximity embedding (SPE). Catalyst and SPE showed the best results.[25]
2008[a]	Conformational sampling of drug-like molecules with MOE and Catalyst: implications for pharmacophore modeling and virtual screening	Evaluation of three conformational search methods of MOE software: systematic search, stochastic search, and conformation import. The performance of each is compared to the others and to Catalyst results. The best parameter sets are empirically determined for optimal conformation coverage and the ability to retrieve the bioactive conformation at different rmsd thresholds. To this end, a set of 256 druglike ligands with their bioactive conformations were selected. However, a sheer quarter of all bioactive structures was not reproduced by Conformation Import at rmsd <1 Å. The tool did not find those molecules with numbers of rotatable bonds and chiral centers superior to the average ones. MOE performed at least as well as Catalyst.[26]
2008	A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS	Field-based methods were compared with each other and with UNITY 2D fingerprints. ROCS including both shape and features along with UNITY fingerprints performed best. Some improvement was seen between using flexible or multiple conformations performed slightly better than rigid 3D approaches.[27]
2008	Molecule-pharmacophore superpositioning and pattern matching in computational drug design	3D pharmacophore approaches and software modeling packages are reviewed. One of the key implications for ligand based pharmacophore is the molecular flexible alignment, because of the conformational flexibility. The correct selection of chemical features allows the pharmacophore model to reflect universal chemical functionalities. Examples of software for pharmacophore modeling are Catalyst, phase, MOE, and LigandScout.[28]

Table 2. continued

pub year	title	key aspects and refs
2008	Targeted rescue of a destabilized mutant of p53 by an in silico screened drug	Rational design of anticancer molecules using structure-based VS. The molecules were experimentally tested, and they bind to destabilized mutants of p53 avoiding its denaturization.[29]
2008	Development of a new pharmacophore model that discriminates active compstatin analogs	Design of a pharmacophore model based on a cocrystallized structure of the third component of complement C3 with an inhibitor and structure activity relationships of nine inhibitors. Six critical structural features for ligand−receptor recognition were identified. The pharmacophore model was tested against a benchmark set, and it was able to identify 70% of the C3 inhibitors, demonstrating to have prediction ability.[30]
2008	How to do an evaluation: pitfalls and traps	Review of certain pose prediction and VS pitfalls. rmsd comparison between experimental and docked poses is questioned because this metric reflects only atomic positions and ignores electron densities or intermolecular interactions. Crystallographic structures are neither precise nor error free due to poor model fitting between electronic density and atomic positions. Enrichment factor is not a perfect metric to assess VS success because it does not allow comparisons among different studies. Biases are present in the selection of decoys to perform retrospective VS.[31]
2008	Evaluation of the performance of 3D virtual screening protocols: rmsd comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes?	Review of certain VS problems and issues. The test set of active and inactive compounds should preferably contain active ones of the same bioassay, and they all should share the same binding mode. The problem with inactive ones is that they are not documented or registered. Therefore, a common practice is to select a set of presumably inactive compounds with similar physicochemical properties to active ones. Input problems appear with the loss of information is caused by format interconversions, incorrect assignation of protonation and tautomeric states. Another issue is the generation of conformations; bias during this step could influence greatly VS results. Target selection is crucial considering an acceptable resolution. Doubts are raised concerning the use of rmsd for molecular comparisons, and enrichment metrics to assess the quality of VS.[32]
2008	Community benchmarks for virtual screening	Introduction of a new public benchmark set of decoys to test docking protocols for VS. Overfitting may occur during docking calculations due to inappropriate selection of the coefficients of a scoring function. Another problem is the bias when selecting decoy ligands for a benchmark set, in focused libraries.[33]
2008	Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase	VS for potential anti-influenza compounds directed to neuraminidase takes into account protein flexibility. Newly designed compounds were docked against a representative ensemble of neuraminidases, and the hits selected were redocked for refinement. Twenty seven compounds were predicted to have a binding affinity similar to known neuramidase inhibitors.[34]
2009	A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor	Shape based VS with ROCS and Topomers outperformed 2D fingerprints for this target and data set. These methods retrieved the most number of actives in retrospective experiments and also the most number of different scaffolds, however, electrostatic similarity methods identified more diverse scaffolds. VS was used in a prospective experiment which identified several chemically different active hits.[35]
2009	Combining docking with pharmacophore filtering for improved virtual screening	Presentation of a new method to overcome the problem of false positives during automated docking for VS. The method consists in docking without scoring after a postfilter process with receptor-based pharmacophores. The results show that the method perform well to eliminate false positives.[36]
2009	The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening	Effects of different tautomer and protonation state are investigated in terms of the success for structure based VS using Autodock. Although the focus is on structure based VS, it is likely that such concerns apply to ligand-based VS.[37]
2009[a]	Conformational sampling with stochastic proximity embedding and self-organizing superimposition: establishing reasonable parameters for their practical use	Evaluation of the following conformational search methods: stochastic proximity embedding (SPE), self-organizing superimposition (SOS), MOE conformational import, MOE bond rotation, and MOE stochastic search. Key objectives were to (i) determine the best-performing parameters—like the maximum number of conformations—for each method to generate a representative conformational sample and (ii) reproduce the bioactive conformation. A data set of 3D drug-like ligands and some more flexible molecules was selected, and different conformational generation protocols were applied. To determine if the bioactive conformation was retrieved, the best fitting conformation of each ligand was clustered by rmsd. SPE and SOS produce satisfactory results with 500 conformations per ligand, whereas MOE conformational search procedures require a higher number of conformations.[38]
2009	Virtual screening of bioassay data	Discussion of three data problems: (i) limited access to curated data; (ii) false positives; (iii) and imbalanced proportion between active and inactive compounds. PubChem is one of the major free accessible resources for bioassay data. However it could have potential errors (for example, false positives) because lacking confirmatory tests for all data. To try to overcome these difficulties there are classification tools like Weka cost-sensitive classifiers, for both primary and comparative studies, that performed relatively well. For the VS of bioassay data, it is recommended that both primary and the corresponding confirmatory screening data are used.[39]
2009	How similar are similarity searching methods? A principal component analysis of molecular descriptor space	It is well-known that different VS methods capture different properties. This paper addresses the question of quantifying the differences between VS techniques and identifying orthogonal approaches. These are important considerations as the field moves toward combining different methods via data fusion.[40]
2010[a]	Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble	For evaluation of the program NMRCLUST to cluster conformational ensembles of drug-like molecules, NMRCLUST is implemented in Chimera. It reduces the number of conformers generated during a conformational search procedure upon selecting the most appropriate structures that will be used for further analysis with the objective of data reduction. The evaluation was done on 65 ligands, using as a reference their bioactive conformations. The ligands were selected and extracted from the PDB. An invaluable asset of this method is that it does not require user-defined rmsd cut off values because it takes into account each conformational ensemble individually.[41]
2010	Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods	A recent study which shows that for retrospective analysis of multiple targets from the DUD (directory of useful decoys) data set, 2D VS methods perform better than 3D methods.[42]
2008	Application of belief theory to similarity data fusion for use in analog searching and lead hopping	This study along with the example for ligand based VS[43] describes the application of belief theory to quantify the probability of finding active compounds by specific VS methods.[43]
2011	A unified, probabilistic framework for structure- and ligand-based virtual screening	This approach could help to identify which methods are most suitable for a given target, or even whether it could be worthwhile performing biological testing of the results from VS.[44]

[a] Literature which refers to benchmarking conformational sampling methods.

Table 3. Issues in VS[a]

pub year	title and refs
2004	Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography[45]
2006	Docking and scoring—theoretically easy, practically impossible?[12]
2008	How to do an evaluation: pitfalls and traps.[31]
2009	How to recognize and workaround pitfalls in QSAR studies: a critical review.[11]
2009	Docking screens: right for the right reasons?[46]
2010	Quo vadis, virtual screening? A comprehensive survey of prospective applications.[8]
2010	How similar are those molecules after all? Use two descriptors and you will have three different answers.[47]

[a] Recommended original publications to expand the reader's knowledge about VS techniques and raise awareness concerning problems with VS.

A straightforward yet very valuable way to assess the relative importance of each individual pharmacophoric feature has recently been described.[30] By eliminating one feature at a time, the authors were able to determine the decrease in hit rate compared to that obtained using the full pharmacophore, and hence, the relevance of each individual feature. In a far more general way, not only pharmacophore queries but any setup condition for which user definitions are provided by the screen algorithm is susceptible to this trade-off between strict and loose search criteria.

c. Pitfall: Predicting the Wrong Binding Pose. Docking-based VS is one of the most frequently used structure-based VS approaches nowadays.[8] However, a docking screen can fortuitously produce the right result for the wrong reasons; that is, correctly assign high scores to many true hits but still predict the wrong binding poses. Hence, some celebrated success stories may simply be the result of serendipity. A recent review of 15 years of research in protein–ligand docking revealed that only few studies present actual evidence that the predicted binding modes were the correct ones.[46] Other reviews even went as far as to conclude that "for prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity".[48]

d. Pitfall: Variable Water-Mediated Binding Interactions. Hydrogen bonds between ligand and protein that are mediated by water are well-documented and are often visible in the crystal structure of the complex. Those water-mediated hydrogen bonds can be taken into account in a structure-based VS study, but it is very difficult to predict the exact number, position, and orientation of these interactions. In this case, the apparent increase in the sophistication of the model to make it resemble the "true" binding environment also becomes a source of uncertainty due to the much greater number of possible states that must be considered, not to mention the substantially greater computational cost.

e. Pitfall: Single versus Multiple/Allosteric Binding Pockets. Both structure- and ligand-based VS approaches have the intrinsic shortcoming that they are unable to identify bioactive ligands for binding pockets which are not explicitly docked against or implicitly represented in the training set. In the experimental validation, ligands may very well (and often do) bind to alternative binding pockets, so the question is whether the obtained enrichment is due to design or serendipity. The fact that the binding site of a ligand is often unknown (such as in radio-ligand binding assays) complicates the problem of properly assessing hit rates in VS experiments.

f. Pitfall: Subjectivity of Post-VS Compound Selection. In practice, when a virtual screen is performed, the virtual hits are reviewed by the project team and a subset of them is selected for physical screening. Our experience shows that certain compromises and the potential for subjective decision making come into play at this point (see also the pitfall regarding the relative weight of substructural features below). Consider the following scenario: a ligand-based VS is performed on an in-house compound collection of 500 000 molecules which identifies the 1000 top-ranking hits. However, the biological screen is performed in a low-throughput mode where there is only capacity to screen 100 molecules, and the compounds need to be cherry-picked by hand. What would be the correct way to proceed in this situation? A simple selection of the top-ranked 100 structures is likely to pick a set of close analogues that may be structurally redundant. A better approach might be to cluster the compounds into related families and select a small number of molecules from each cluster (or scaffold). While this approach is perfectly sensible, it introduces a bias (subjectivity) in the VS protocol, rendering comparisons difficult.

g. Pitfall: Prospective Validation. The benchmarking of VS algorithms is usually performed on data sets with known active but often only putatively inactive molecules. Once the best-suited VS protocol(s) are identified and published, they are rarely discussed and externally validated in a prospective context. In contrast, QSAR studies demand rigorous validation of the models including experimental verification on new data sets not considered during model development.[11] This is not a common practice in VS, although some journals now started implementing editorial guidelines to that effect (such as the *Journal of Medicinal Chemistry*).

h. Pitfall: Drug-likeness. Some VS approaches (either on the level of study design or software algorithm) are based on "drug-like" compounds, as defined by Lipinski in his "rule-of-five" work.[49] However, it should be kept in mind that this only applies to oral bioavailability and that many bioactivity classes such as antibiotics routinely fall outside the scope of this rule. Hence, VS protocols are generally applied and validated on a relatively small fraction of chemical space, and their performance may change drastically when venturing outside the charted territory of known drug classes (e.g., when seeking novel compounds for targeting protein–protein interactions which are currently thought to mostly reside out of "Lipinski Space").

i. Pitfall: Diversity of Benchmark Library Vs Diversity of Future, Prospective VS Runs. A large number of published VS studies are performed within the chemical space of public and commercial (typically "drug-like") libraries.[10] The type of screening library however needs to be closely related to the objective of the particular VS campaign, and it needs to be made sure that results are transferable between runs.[50] Chemically diverse libraries are particularly attractive for identifying novel scaffolds for new or relatively unexplored targets, such as those from diversity-oriented synthesis.[51,52] If the goal of the screening is directed at a specific target family, one may use target-oriented synthesis (TOS),[53] focused, or targeted libraries.[54] If the goal is lead optimization, chemical libraries with high intermolecular similarity (e.g., combinatorial libraries)[55] are an attractive source. The structures of natural products are, in general, different from those of synthetic compounds and occupy different areas of chemical space.[56]

Interestingly, Lipinski has commented that a large number of natural products are bioavailable[57] and the rationale of these observations was recently provided.[57] In addition to commercial databases such as the Dictionary of Natural Products,[58] other databases include a subset of ZINC and a collection recently released by Specs.[59] Also a unique collection of natural products can be used in VS through the Drug Discovery Portal.[60] Overall one needs to remember that the library must fit the purpose of the experiment (see also Table 4).

2. PITFALLS CONCERNING DATA DESIGN AND CONTENT

a. Pitfall: (Non)Comparability of Benchmark Data Sets. In practice, it is very difficult to determine why a VS study was successful, particularly if multiple methods have been employed. Sometimes authors evaluate the performance of their VS methods at hand (benchmarking), while another approach is to compare the robustness of benchmarking conclusions in meta-analyses. The importance of the latter can be understood due to the fact that primary literature in many cases employs different target classes and data sets, rendering a meta-level analysis important.[82−84] Overall, it is probably correct to state that in many cases ligand-based (2D-) methods tend to outperform docking methods, and that (2D-) fingerprint-based methods generally outperform (3D-) shape-based approaches, such as in the case of the target proteins contained in the Database of Universal Decoys (DUD).[85,42] In the same spirit, MacGopher and co-workers found that averaged over multiple targets, ligand-based methods outperformed docking algorithms on the data sets studied.[86,21]

The comparative performance of VS is dependent on the query molecule(s), biological target, and/or chemical library being screened.[83,84] With the availability of many different computational approaches, it is good practice to evaluate their performance on a few standard data sets (e.g., DUD[85]) to enable direct comparison of methods.[42,85] For such studies to be meaningful, all parameters must be clearly described. Potential issues were discussed in detail in a special issue of the *Journal of Computer-Aided Molecular Design*[87] and, where possible, best practices (such as those outlined by Rohrer et al.[88]) should be followed. Recent initiatives, such as the “Online Chemical Database with Modeling Environment”,[89] which aim at publishing both the actual models and the data sets used to develop them, can have a positive impact on the ability to share, benchmark and evaluate models in a reproducible manner. A similar tool, Chembench,[90] provides rigorous model validation routines; however in its current version, no provision has been made for storing and sharing data alongside the models developed.

b. Pitfall: Limited Comparability of Performance Metrics. While data sets used for benchmarking VS algorithms are often difficult to compare, the same can also be said about the performance indicators employed. In this context, using mean enrichment factor (EF) as a comparative performance measurement is rather risky, because the values obtained highly depend on the ratio of active to inactive molecules in the database screened.[91] A pitfall concerning EF is the saturation effect (ceiling/plateau of the curve to unit value 1), making the comparison of benchmarking results between VS studies less distinguishable.[82] Other researchers also reported the problem for performance metrics—like enrichment and ROC plots—because their values do not distinguish well in areas of early and late performance (starting and ceiling of the curve).[42] This led to the establishment of measures such as the BedROC score, which attributed different importance to early and late stages of the retrieved list of compounds.[92]

c. Pitfall: Hit Rate in Benchmark Data Sets. Two factors that complicate benchmarking of VS algorithms are the size and diversity of the chemical libraries. In the early days, benchmark libraries were either too small or contained too many closely related analogues—and often both.[47] Small libraries are not representative of most real-world applications, where confirmed hit rates often range between 0.01% and 0.14%.[93] Similarly, libraries that are too homogeneous artificially inflate the performance of the method. This “analogue bias” can be manifested not only at the structural level, but at a coarse physicochemical property level as well.[20,94,95] In order to remedy this problem, recent studies have focused the selection on subsets of active molecules[96,97] or reported enrichment in terms of distinct molecular scaffolds retrieved in the hit list.[40,98] What is less obvious, however, is that it is not only the *active* molecules that are crucial in assessing performance, but the *inactive* ones as well. Consider, by analogy, training a classifier to recognize different means of human transportation from pictures of horses, cars, trains, airplanes, and so on. If the negative data set includes very different items (say, desks, chairs, etc.), then distinguishing human transportation vehicles from other totally unrelated items is relatively simple. However, if the data set also includes trucks for transporting goods or minibuses which can be used either way, the classification task becomes more challenging. The same situation is encountered in assessing VS methods, where the size of the data set and the diversity of the molecules in the active and inactive classes need to be taken into account. Recent approaches such as the one reported in ref 88 addresses this concern by constructing a “maximum unbiased validation” (MUV) data set (see below); however, earlier studies often did not pay sufficient attention to this issue.[47] Data set biases are a frequent source of the observed enrichments, particularly in early comparative VS studies. For ligand-based VS, a recent study[20] found that about half of the VS performance could be obtained by simply using the number of atoms per element as a descriptor.

d. Pitfall: Assay Comparability and Technology. The importance of properly designed data sets for comparing VS algorithms was addressed only recently, most notably in the publication of the MUV concept.[88] In that work, for the first time, the similarity distributions within the active as well as the inactive parts of the data set were considered. In addition, the data was derived from PubChem,[99] which is publicly available and includes all the details about the assay setup which are typically not provided in other databases. The MUV data set is large and excludes some questionable compounds such as those exhibiting autofluorescence, which is why it was chosen for subsequent benchmarking studies.[100] However, even this data set is unlikely to be perfect, such as due to the data source employed. Although PubChem[10,76] is clearly an important source of structure−activity information, the data deposited there have not been reviewed and scrutinized by an independent source, and assays from different origins may not be of the same quality. Some data may be incomplete and there could be cases where the information deposited is simply wrong or inconsistent with other sources. One example of disagreement with previously published results is the hERG assay (PubChem AID 376), which shows little agreement with patch clamp data.[101] While part of this disagreement is likely due to the different assay approaches, the question remains what to do

Table 4. Listing of Available Online Compound Databases for Screening[a]

databases and ref	web pages	number of molecules	formats	free
Binding Database[61]	http://www.bindingdb.org	284 206 small ligands with 648 915 binding data, for 5662 protein targets	SDF	yes
Chem ID[62]	http://chem.sis.nlm.nih.gov/chemidplus/	388 000	access to chemical substances in the National Library of Medicine database (NLM) through identification files	yes
ChemBank[63]	http://chembank.broadinstitute.org	800 000	SDF, Txt, SMILES	yes
ChEMBL db[64]	https://www.ebi.ac.uk/chembldb/	658 075 differing bioactive compounds and 8091 targets	SDF, Txt, SMILES, FASTA	yes
Chemical Entities of Biological Interest (ChEBI)[65]	http://www.ebi.ac.uk/chebi/init.do	584 456	SDF, OBO, Flat File, Oracle Binary table, dumps, Generic SQL	yes
ChemMine[66]	http://bioweb.ucr.edu/ChemMineV2/	6 200 000	SDF, SMILES	yes
Chimiotheque nationale[67]	http://chimiotheque-nationale.enscm.fr/index.php	44 817 compounds, 32 573 compounds in plate, 14 514 natural extracts	SD	yes
Commercial Compound Collection (CoCoCo)[68]	http://cococo.unimore.it/tiki-index.php	6 957 134 molecules, more than 144 millions conformations	SDF	yes
Developmental Therapeutics Program (DTP)[69]	http://dtp.nci.nih.gov Downloads: http://cactus.nci.nih.gov/	473 965	XML, html, SDF	yes
DrugBank[70]	http://www.drugbank.ca	6827 drugs, 4477 nonredundant protein sequences	SDF, DrugCard Format, FASTA	yes
GVK BIO[71]	http://www.gvkbio.com/informatics.html	not specified (focused libraries with target inhibitor or toxicity collections applied in the field of bio- and chemo-informatics)	ISIS/Base DB, SD, XML and Oracle format among others not specified.	no
i:lib diverse[72]	http://www.inteligand.com/	drug-like fragment set for combinatorial library generation	SDF, SMILES	no
Mother of All Databases (MOAD)[73]	http://www.bindingmoad.org	14 720 ligand-protein complexes, 4782 structures with binding data, 7064 ligands	CSV, PDB, SMILES	yes
PDB bind[74,75]	http://www.pdbbind.org/	3214 ligand−protein complexes	PDB, Mol2, SD	yes[b]
PubChem[76]	http://pubchem.ncbi.nlm.nih.gov/	49 875 000	Text ASN.1, Binary ASN.1, XML, SDF, Image/Small image, SMILES, InChI	yes
Therapeutic Target Database[77,78]	http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp	1906 targets, 5124 drugs	SDF, Mol, FASTA	yes
Traditional Chinese Medicine Database (TCM)[79]	http://tcm.cmu.edu.tw/	more than 20 000 pure compounds isolated from 453 TCM ingredients	CDX, Mol2	yes
WOMBAT[80]	http://www.sunsetmolecular.com/	305 727 molecules, 1966 unique targets	CSV, RDF, SMILES	yes[c]
ZINC[81]	http://zinc.docking.org/	13 000 000	SMILES, Mol2, SDF, pdbqt, flexibase formats	yes

[a] The second to last column indicates possible free download offers. (Databases were last visited in February 2011.) The listing complements and updates an earlier recompilation.[10] [b] For academic sites only. [c] Free samples only.

(how to combine and how to compare) with bioactivity data from different sources and what to make of the disparate results obtained. A detailed analysis of which data source is more trustworthy is beyond the scope of this review, the case highlights the general problem of ambiguities arising from the misuse of public databases which themselves do not require careful curation and quality checks on the data that is being deposited.

Another unavoidable source of disagreement when assembling VS benchmark data sets relates to the choices of particular targets, molecular weight, and other property cut-offs, and similarity/diversity measures and selection algorithms used to assemble the screening library. Many of these parameters are completely subjective (such as the definition of what constitutes a distinct chemotype or cluster), so it is unrealistic to expect that a universally accepted "ideal" VS benchmark deck will ever emerge.

e. Pitfall: "Bad" (Reactive/Aggregating etc.) Molecules. Compound collections provided for VS (such as vendor libraries) often include molecules that contain chemically reactive groups or other undesirable functionalities that interfere with the HTS detection techniques and cause them to elicit a positive signal. For example, some compounds may be autofluorescent and others may aggregate at certain concentrations and produce a spurious response.[102] In short, these "bad" molecules encompass chemically reactive, assay-interfering compounds, and are often referred to as PAINS (pan-assay interfering substances) or frequent hitters.[103] There are also compounds that interfere with the readout of a particular assay type, such as kinase inhibitors that cause false positive readouts in reporter-gene assays.[104] Hence, it is likely that a substantial number of molecules reported to be bioactive against a set of proteins are false positives, which is an even more problematic situation than the reverse (missing compounds which are actually active or false negatives) due to the smaller number of active compared to inactive compounds in training sets.

f. Pitfall: Putative Inactive Compounds As Decoys. Known (experimentally confirmed) inactive compounds are useful as negative controls because only few of them should appear in the hit list when a reliable VS protocol is employed. However, many of the decoys used in VS benchmark studies are only putatively inactive; hence, some assumed true negatives may actually be positives in reality. This problem becomes even more severe when the negative decoys resemble the true actives not only in a physicochemical sense but also chemically, since this substantially increases the chance that the putatively inactive molecules may be active against the target under investigation.

The pros and cons of library enrichment with four different types of decoys for benchmarking tests were exhaustively discussed by Nicholls in 2008.[83] Randomly sampling universal decoys falls short of representativeness facing the sheer number of molecules in chemical space (an estimated 10^{56}) and due to "inductive bias" of the sampler (cf. expectation bias).[11] Another reported pitfall, related to the topic of this paragraph, denoted "false false positives", which arises when an actually active compound has not been annotated as such and is predicted to be active at the same time.[83] In this regard, collecting drug-like decoys has the advantage "that because decoys are derived from characterized collections they are more likely to be known to be inactives. This holds only if decoys are selected for their drug-likeness combining similar physicochemical similarities with chemical dissimilarities what makes them improbable to share the same target".[83]

g. Pitfall: Feature Weights. Ligand-based VS, based on a single query, typically places equal importance on all parts of the molecule. However, some substructural features may not be required for activity against a specific target of interest. Hence, where possible, the query should contain the minimum number of features that are believed to be relevant for activity. Pharmacophore queries, as well as fingerprint-based bioactivity models, can be tuned to circumvent this problem if they are built from multiple active molecules and submitted to rigorous statistical feature selection prior to performing the virtual screen. In addition, newer versions of 3D shape-based VS tools, such as Rapid Overlay of Chemical Structures (ROCS),[105] allow editing of the query to remove certain features that are not deemed critical for activity.

3. PITFALLS RELATING TO THE CHOICE OF SOFTWARE

a. Pitfall: I/O Errors and Format Incompatibilities. As pointed out by Kirchmair and Langer et al.,[32] a more mundane but serious problem are errors introduced when interconverting different molecular formats. Because molecular modeling represents a niche market and not all pieces of software are subject to the same quality control standards, it is often the case that information may get lost or altered when converting one file format to another, or even when using the same format in different pieces of software. The information that can get distorted ranges from benign annotations, to more serious issues, such as atomic coordinates, chirality, hybridization, and protonation states, etc. (see also Table 4).

b. Pitfall: Molecule Preparation. Adding implicit hydrogen atoms and assigning the correct charges can easily be forgotten in many VS algorithms that depend critically on these parameters. This is not a trivial problem, particularly when multiple software tools are employed in a VS pipeline, because some programs expect the user to explicitly handle hydrogen atoms, charges, ionization states, and so forth, while others do this type of preprocessing in a fully unsupervised manner.[32] As an illustration, certain modeling tools like Sybyl[106] are unable to assign formal charges in an automated way when calculating partial charges, whereas other programs like Vega ZZ[107] perform this step on-the-fly when partial charges are requested. In MOE pharmacophore searching, the user is required to assign the appropriate ionization states prior to automated partial charge calculation. Catalyst does not rely on charges and allows relevant moieties to be treated as ionizable.[32] Indeed, different tools differ significantly in the degree of control they give to the user over the database and query preparation process, a critical step that is sometimes forgotten or paid little attention to. As a general rule, the query molecule(s) must be preprocessed exactly the same way as the structures in the database being screened to ensure consistency.

c. Pitfall: Feature Definition. In pharmacophore queries, the definition of pharmacophore features needs to be applied with caution. For example, it is known from crystallographic evidence that nitrogen and oxygen atoms in the same heterocycle such as an oxazole do not both behave as hydrogen bond acceptors (HBA) simultaneously.[108] In the overwhelming majority of cases, the acceptor in the oxazole ring is the nitrogen. There are many other cases where oxygen does not behave as an HBA, such as the ether oxygen in an ester, the oxygen of a furan ring, etc. However, not all computational methods take that context into account. There are at least three areas of context-dependent chemical characteristics that are relevant to VS: (1) tautomeric form, where the selection of the wrong tautomer can misguide the assignment of hydrogen bond acceptors or donors, leading to false positives, and/or false negatives; (2) ionization, where the protonation state of chemical groups at physiologically relevant pH can be miscalculated, often because of the inability to predict whether a particular acidic or basic moiety is charged or uncharged in the hydrophobic interior of a protein; and (3) chirality, where for racemic structures one needs to calculate the conformations for all possible chiral configurations.[32]

d. Pitfall: Fingerprint Selection and Algorithmic Implementation. In similarity-based screening, performance depends critically on the choice of descriptors. A comparison of the behavior of different descriptors in retrieving molecules from a database has recently been published.[40] In this work, four broad descriptor classes were considered, namely circular fingerprints, circular fingerprints with counts, path-based and keyed fingerprints, and pharmacophore descriptors. The authors concluded that descriptor behavior was much more determined by those four classes than the specific parametrization. In a related unpublished work carried out by one of the authors of the present manuscript, predefined MACCS keys (which are based on precalculated molecular fragments) were compared with typed graph distances (TGD, which capture atom-centric interactions such as hydrogen bond donors and acceptors) and showed substantially different performance against the training and test sets. These are just two recent examples from a very rich literature on this topic.[98,109−113] It is also important to remember that some fingerprint types, such as MACCS keys, are only partially disclosed publicly, leading to different implementations in different packages (see ref 114 for a recent discussion on this topic). Also, as with all software, descriptors may be implemented in a different way within different software packages, though averaged over large databases this effect seems to be less extreme than in other cases.[40]

e. Pitfall: Partial Charges. It may happen that a particular pharmacophore feature requires a localized full or partial charge on a specific atom or site, but pin-pointing that charge is difficult due to resonance. As an illustration, it is not possible to obtain the positively charged nitrogen atom on a monocationic guanidinium fragment calculated *via* Gasteiger charges, since the formal +1 charge is redistributed through the guanidine structure. Even worse, it could happen that the central carbon atom is assigned a formal charge of +1, which would be chemically and biologically meaningless. The obvious workaround is to manually edit the default charge assignments, but this is impractical for large libraries which require fully automated molecule preprocessing. In cases where mesomeric cation and anion centers, such as cationic ammonium from either lysine or guanidinium and anionic carboxylate of aspartic acid or glutamic acid, form salt bridges, it may be more prudent not to represent the mesomeric system at all but rather assign full charges of +1 and −1 to the corresponding atoms. Such localized full charge representations are more biologically meaningful and expected by some VS software such as MOE, where cation and anion features are predefined and fixed on distinct atoms like N or O to define pharmacophore patterns for 3D screens.

f. Pitfall: Single Predictors versus Ensembles. An often experienced and therefore commonplace experience for the VS practitioner is that different screen methods retrieve different molecules from the same database. Thus, it is a common practice to run several VS methods for any given target and additively collect the outcome.[115] One of the safest conclusions drawn from previous VS comparative studies is that no single method works best in all cases. Since the best method is not known *a priori*, a recommended practice is to use multiple methods and combine their results (consensus study). This approach has been used for many years in the field of statistical learning. When the outcome from several slightly different input subsets are "bagged into a common result bag", we speak of bagging.[115] The ensemble of data fusion techniques, such as boosting[116] and stacking,[115] combine multiple models to achieve better predictive performance than could be obtained from any of the constituent models. Obviously, combining the output of multiple predictors is useful only if there is disagreement between them, so much of the work in this field has been devoted to methods for introducing diversity into the model pool.[40] Ensemble techniques are becoming increasingly prevalent in QSAR-based,[117−119] similarity-based,[120−123] and structure-based virtual screening.[124,125] Different aggregation techniques for merging the results of multiple methods are discussed by Willett.[126,127]

4. PITFALLS CONCERNING CONFORMATIONAL SAMPLING AS WELL AS LIGAND AND TARGET FLEXIBILITY

a. Pitfall: Conformational Coverage. One of the major challenges in 3D VS is generating a manageable set of conformations that adequately cover the molecule's conformational space. To achieve this goal, several key parameters need to be empirically tested and optimized, including: (1) sampling algorithms and their specific parameters; (2) strain energy cutoffs; (3) maximum number of conformations per molecule; and (4) rmsd thresholds used to remove duplicates. The training set must obviously contain several control molecules whose bioactive conformations are known, most commonly from X-ray cocrystals or NMR studies. Following good modeling practices, the best parameters identified from the training set must also perform well against an independent test set; that is, only when the method successfully retrieves bioactive conformations of other control molecules which are not part of the training set can one assume that the effectiveness and reliability of the established protocol is sufficiently proven and thereby positively validated. Several publications have tackled this topic,[13,19,23,25,26,38,41] and some methods have been specifically tuned to increase the odds of identifying bioactive conformations.[128] However, the enormous size of conformational space accessible to many molecules of pharmaceutical interest (a molecule with six rotatable bonds sampled in 10° intervals will give rise to $36^6 \approx 2 \times 10^{10}$ conformations) makes such comparisons a formidable challenge in practice.

b. Pitfall: Defining Bioactive Conformations. Good sampling is critical because our knowledge of pharmacologically relevant conformational space is very limited. Reproducing known ligand geometries is insufficient because these represent an extremely limited and biased sampling of all bound ligand conformations. Most ligands have never been cocrystallized with their primary targets, and even fewer have been cocrystallized with important countertargets. Hence, the same ligand may bind to different proteins in vastly different conformations, depending on the geometry and flexibility of the binding site. Most importantly, crystal structures are not obtained under physiological conditions, which casts doubts on whether the conformations of the cocrystallized ligands are indeed the bioactive ones.[45] Experience gathered from protein−ligand structures, NMR analyses of solutes, and computational studies suggests that solution structures of small organic compounds and their active site conformations are not the same at all. In addition, large differences exist between the conformation of a ligand seen in a crystal protein complex and that of the free molecule. It is now widely accepted that upon binding to a protein, ligands can undergo substantial conformational changes.[129]

Special tools for alignment-free three-dimensional VS seem not to require a conformational ensemble, based on the observation that one conformation alone, which does not necessarily need to be the bioactive one either, resulted in a significant enrichment of the hit list (here applied to "three-dimensional similarity searching with pharmacophore-based correlation vectors").[130] Typically, the "bioactive" or "receptor-relevant" conformation remains unknown unless it has been cocrystallized in its target complex. Minimum energy conformations are a common surrogate. However, a study shows that a "consensus" conformation (highest average overlap with all conformations) outperforms those minimum ones during VS.[131] Wolber and co-workers discuss the downside of entirely shape-based VS in compensating the loss of physicochemical and pharmacophore information through enhanced queries.[91]

c. Pitfall: Comparing Conformations. Even assuming that the bioactive conformation is known, the question is raised as to what constitutes a hit. In other words, how similar a computationally generated conformation must be to the bioactive one to be declared a match? The most common measure for determining geometric deviation from a given reference structure is the rmsd, with most publications considering values less than 0.25−0.3 Å sufficient to call two conformations "identical".[13] Other authors suggest that for small organic molecules two conformations can be considered identical if their rmsd is less than 0.1.[19] According to the same group, rmsd values between 0.1 and 0.5 represent an excellent fit, between 0.5 and 1.0 a good fit, between 1.0 and 1.5 an acceptable fit, between 1.5 and 2.0 a less acceptable fit, and above 2 not a fit at all, at least in a biological context.[19,38] Moreover, rmsd values reported in the literature may not be directly comparable, given that they may have been derived using different superpositioning algorithms and may have been based on all atoms or just certain fragments.[28,32] A different method for determining conformational similarity is to compare the actual electron densities.[31] This approach was motivated by the fact that most conformational sampling algorithms are evaluated against experimental structures derived from X-ray crystallography, where the actual observable is the electron density itself and not the atomic positions; the latter are derived through an optimization process that refines the atomic coordinates so as to maximize the fit to the observed electron density. By directly comparing electron densities, the authors argue that one avoids the risk of using an intermediate structure that could be wrong (i.e., does not fit the experimental density very well), as is the case with a significant number of structures deposited in the PDB.

As mentioned above for alignment-free approaches, it is not clear that having the bioactive conformation present in the query and database molecules is a prerequisite for having a successful VS campaign. This was also the case for shape-based VS using ROCS; a recent study[132] showed that the simplest settings (even using a single conformation as a query) often yielded the best, or among the best, results, thus allowing efficient VS to be performed against multiple bioactivity classes. Of course, the main concern with such observations is their statistical strength and generalizability. A situation where the right answer is produced for the wrong reason is always a good cause for skepticism and further investigation.

d. Pitfall: Size of Conformational Ensemble. Of course, one should not expect that every conformation generator is capable of producing the bioactive conformation of interest. Therefore, a practical question that is often asked is how many conformations need to be calculated to have sufficient confidence that the bioactive one is included in the resulting ensemble. Ideally, one would wish to have as many conformations as possible for each molecule to ensure adequate coverage. However, a thorough conformational search for millions of database molecules would require enormous computational resources, which are typically not available. Therefore, one needs to find a compromise between computational cost and sampling breadth. But no matter how much computational power one chooses to apply to the problem, in our view, being able to retrieve bioactive conformations is a reassuring feature, but should not be the sovereign measure for assessing the utility of a conformational sampling algorithm in the context of VS. Recovering a known set of bioactive conformations does not guarantee that other bioactive conformations of the same ligands will be present in the ensemble.

e. Pitfall: Ligand Flexibility. A common practice in many 3D database search systems is to set a limit on the number of conformations stored for each molecule. As stated above, that number is chosen empirically to achieve a compromise between accuracy (i.e., rmsd to biologically active conformer) on the one hand, and database size, (i.e., the number of conformers in an ensemble) with its implication on storage space and computational speed on the other hand. The number of conformations accessible to a molecule depends greatly on its size and flexibility. Griewel and colleagues demonstrated that for molecules with less than nine rotatable bonds, ensembles with an average accuracy better than 1 Å can be generated with some twenty conformers; however, significantly more conformers are needed for more flexible molecules.[133] If there are only a handful of rotatable bonds in every compound in the database, setting a uniform limit can be suitable. In a recent study, five conformations have been used and were shown to yield high enrichment scores, with the authors pointing out that their results are target protein dependent.[84] However, the situation is very different when the database contains molecules that range wildly in flexibility or topology. The latter is a particularly challenging problem, because some conformational search methods work well only for certain types of molecules. For example, very few of the available methods can deal effectively with constrained systems such as macrocycles, which are becoming increasingly attractive as potential drugs because they are less encumbered by intellectual property constraints. Two closely related methods that have shown great promise in this regard are stochastic proximity embedding (SPE) and self-organizing superposition (SOS).[134] Since the conformations used to populate a 3D database are usually constructed by the same method, it is important that this method can handle well and consistently all possible types of molecules. Regardless of what method is used, one needs to select a number of representative conformations from a much larger pool, a task that can be accomplished through clustering.[41]

f. Pitfall: High Energy Conformations. While good conformational coverage is very important, high energy or physically unrealistic conformations can be detrimental to VS. Some conformational sampling methods do not employ energy minimization to refine and properly rank the resulting geometries, and as a result high-energy conformations can make it into the final ensemble. If such geometries are not eliminated either when the database is constructed or when the virtual hits are identified, the resulting list could contain many false positives. This problem manifests itself often in 3D pharmacophore searches where high-energy conformers which satisfy the 3D arrangement of the query features can be retrieved. As reported before, about 70% of ligands bind at strain energies below 3 kcal/mol,[135] which is a sensible, albeit stringent, threshold to use for filtering out irrelevant conformations.

g. Target Flexibility. Of course, it is not only the ligands that are flexible; it is the biological targets as well. Protein flexibility is probably the most unexploited aspect of VS, mainly because of the computational cost and complexity required to properly model it. The time scale of protein motions range from femtoseconds to seconds, and each of them provides important and specific information.[136] Although protein flexibility is implemented in several docking software, it is rarely used in VS, where the number of molecules that need to be screened is very large. Another issue is failing to capture important aspects of the thermodynamics governing target recognition, such as in cases where the target interaction is not well understood (i.e., where the flexibility of a binding pocket is not taken into account) or when the software does not offer the appropriate algorithms.[15,30]

h. Pitfall: Assumption of Ligand Overlap. In 3D shape-based VS, most programs attempt to maximize the overlap between the query and the database molecules. However, we know from X-ray structures of different molecules in the active site of the same protein that this overlap is sometimes not very large. Indeed, different ligands may occupy different regions in the same protein, even in the same binding site, and the overlap between them in 3D space can be much less than assumed by a shape-based VS tool, resulting in more false negatives. While docking can identify molecules which bind to an active site in very different ways, this is usually not the case for ligand-based VS.

i. Pitfall: Missing the Positive Controls. This pitfall relates to the inability to retrieve known positive controls in a VS run due to applying too strict a cutoff in the first place—this issue is most prominent when applying a very strict cutoff in pharmacophore models. Hence, the selectivity/sensitivity cutoff needs to be determined appropriately when determining parameters of this type, in order to achieve proper benchmark statistics when calculating hit rates in virtual screening runs.

CONCLUSIONS

This review summarized a set of pitfalls that are commonly encountered in VS. Just like physical screening, VS can be executed in a fully automated and largely unsupervised mode. Equipped with the right software, a practitioner can execute a VS campaign against any conceivable library and any conceivable target almost effortlessly—which represents an opportunity, but also a danger. As we hope to have demonstrated, VS requires careful preparation of the database, judicious parameter choices, and sensible compromises between the different goals one attempts to obtain.

Benchmarking studies is a good way to identify algorithms and parameter settings with broad utility. The value of these studies and the confidence in the reported outcomes can be greatly enhanced through public availability of data and easy access to the underlying algorithms. Through a series of government-funded initiatives which mandate deposition of screening data in public repositories and promote the use of open source software, today's data sets are larger than ever before, span more bioactivity classes, and address many of the biases inherent in earlier benchmark collections. While VS is and will always remain a probabilistic game, the odds of success can be greatly enhanced through careful planning and attention to detail. After all, chance always favors the prepared mind.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tscior@gmail.com. Phone: +52-222-2-295500 ext 7529. Fax: +52-222-2-295584.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to BUAP/VIEP for supporting the project SCJT-NAT 10 I 2010 and 2011 and for financial support by CONACyT for Scholarship 230292 to K.C.C. A.B. thanks Unilever for support. J.L.M.-F. and K. M.M. are grateful to the State of Florida for funding.

LIST OF ABBREVIATIONS AND GLOSSARY

2D/3D, two-/three-dimensional; benchmarking, performance tests of different VS methods and evaluation of their findings (retrospective VS experiments); decoy, dummy congener, an inactive imitation (false positive) of an active molecule (true positive); false positive, inactive compound is recognized as active (true positive) during VS; FP, fingerprint; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; hERG, human ether-à-go-go related gene; HTS, high-throughput screening; LB, ligand-based; logP, logarithm of the partition ratio of the concentrations of the un-ionized solute in water against octanol solvent; MACCS, molecular ACCess system (and a fingerprint consisting of 166 predefined keys); MDB, molecular database(s); MUV, maximum unbiased validation; MW, molecular weight; NMR, nuclear magnetic resonance; PB, protein-based; PDB, protein data bank; (Q)SAR, (quantitative) structure–activity relationships; rmsd, root mean square deviation; TGD, typed graph distance; TOS, target-oriented synthesis; VLS, virtual library screening, here synonym to VS; VS, virtual screening

REFERENCES

(1) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Predicting Molecular Interactions in Silico: I. A Guide to Pharmacophore Identification and Its Applications to Drug Design. *Curr. Med. Chem.* **2004**, *11*, 71–90.

(2) Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal Assignment Methods for Ligand-Based Virtual Screening. *J. Cheminform.* **2009**, *1*, 14.

(3) Villoutreix, B. O.; Renault, N.; Lagorce, D.; Sperandio, O.; Montes, M.; Miteva, M. A. Free Resources to Assist Structure-Based Virtual Ligand Screening Experiments. *Curr. Protein Pept. Sci.* **2007**, *8*, 381–411.

(4) Ekins, S.; Mestres, J.; Testa, B. In Silico Pharmacology for Drug Discovery: Applications to Targets and Beyond. *Br. J. Pharmacol.* **2007**, *152*, 21−37.

(5) Guido, R. V. C.; Oliva, G.; Andricopulo, A. D. Virtual Screening and Its Integration with Modern Drug Design Technologies. *Curr. Med. Chem.* **2008**, *15*, 37−46.

(6) Seifert, M. H. J.; Lang, M. Essential Factors for Successful Virtual Screening. *Mini-Rev. Med. Chem.* **2008**, *8*, 63−72.

(7) Waszkowycz, B. Towards Improving Compound Selection in Structure-Based Virtual Screening. *Drug Discovery Today* **2008**, *13*, 219−226.

(8) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.

(9) Böhm, H.; Schneider, G. *Virtual Screening for Bioactive Molecules*; WILEY-VCH Verlag GmbH.

(10) Scior, T.; Bernard, P.; Medina-Franco, J. L.; Maggiora, G. M. Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery. *Mini-Rev. Med. Chem.* **2007**, *7*, 851−860.

(11) Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297−4313.

(12) Coupez, B.; Lewis, R. A. Docking and Scoring—Theoretically Easy, Practically Impossible? *Curr. Med. Chem.* **2006**, *13*, 2995−3003.

(13) Boström, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.

(14) Diller, D. J.; Merz, K. M. Can We Separate Active from Inactive Conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105−112.

(15) Steindl, T.; Langer, T. Influenza Virus Neuraminidase Inhibitors: Generation and Comparison of Structure-Based and Common Feature Pharmacophore Hypotheses and Their Application in Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1849−1856.

(16) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862−865.

(17) Steindl, T.; Langer, T. Docking Versus Pharmacophore Model Generation: A Comparison of High-Throughput the Search of Human Rhinovirus Virtual Screening Strategies for Coat Protein Inhibitors. *QSAR Comb. Sci.* **2005**, *24*, 470−479.

(18) Langer, T.; Wolber, G. Virtual Combinatorial Chemistry and in Silico Screening: Efficient Tools for Lead Structure Discovery? *Pure Appl. Chem.* **2004**, *76*, 991−996.

(19) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.

(20) Bender, A.; Glen, R. C. Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.

(21) Knox, A. J. S.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in Compound Database Preparation-"Hidden" Impact on Virtual Screening Results. *J. Chem. Inf. Model.* **2005**, *45*, 1908−1919.

(22) Wolber, G.; Langer, T. Ligandscout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(23) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.

(24) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(25) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007**, *47*, 1067−1086.

(26) Chen, I. J.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773−1791.

(27) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719−729.

(28) 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.

(29) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted Rescue of a Destabilized Mutant of P53 by an in Silico Screened Drug. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10360−10365.

(30) Chiu, T.-L.; Mulakala, C.; Lambris, J. D.; Kaznessis, Y. N. Development of a New Pharmacophore Model That Discriminates Active Compstatin Analogs. *Chem. Biol. Drug Des.* **2008**, *72*, 249−256.

(31) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−190.

(32) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the Performance of 3D Virtual Screening Protocols: RMSD Comparisons, Enrichment Assessments, and Decoy Selection - What Can We Learn from Earlier Mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213−228.

(33) Irwin, J. J. Community Benchmarks for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193−199.

(34) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-Based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878−3894.

(35) Tresadern, G.; Bemporad, D.; Howe, T. A Comparison of Ligand Based Virtual Screening Methods and Application to Corticotropin Releasing Factor 1 Receptor. *J. Mol. Graphics Modell.* **2009**, *27*, 860−870.

(36) Peach, M. L.; Nicklaus, M. C. Combining Docking with Pharmacophore Filtering for Improved Virtual Screening. *J. Cheminf.* **2009**, *1*, 6.

(37) Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742−2748.

(38) Tresadern, G.; Agrafiotis, D. K. Conformational Sampling with Stochastic Proximity Embedding and Self-Organizing Superimposition: Establishing Reasonable Parameters for Their Practical Use. *J. Chem. Inf. Model.* **2009**, *49*, 2786−2800.

(39) Schierz, A. C. Virtual Screening of Bioassay Data. *J. Cheminf.* **2009**, *1*, 21.

(40) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.

(41) Yongye, A. B.; Bender, A.; Martinez-Mayorga, K. Dynamic Clustering Threshold Reduces Conformer Ensemble Size While Maintaining a Biologically Relevant Ensemble. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 675−686.

(42) Venkatraman, V.; Perez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools against the DuD Data Set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079−2093.

(43) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941−948.

(44) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Merta, P.; Locklear, J.; Hajduk, P. J. A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening. *J. Med. Chem.* **2011**, *54*, 1223−1232.

(45) DePristo, M. A.; de Bakker, P. I. W.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure* **2004**, *12*, 831−838.

(46) Kolb, P.; Irwin, J. J. Docking Screens: Right for the Right Reasons? *Curr. Top. Med. Chem.* **2009**, *9*, 755−770.

(47) Bender, A. How Similar Are Those Molecules after All? Use Two Descriptors and You Will Have Three Different Answers. *Expert Opin. Drug Discovery* **2010**, *5*, 1141−1151.

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

(49) 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.

(50) Shelat, A. A.; Guy, R. K. The Interdependence between Screening Methods and Screening Libraries. *Curr. Opin. Chem. Biol.* **2007**, *11*, 244−251.

(51) Spandl, R. J.; Bender, A.; Spring, D. R. Diversity-Oriented Synthesis; a Spectrum of Approaches and Results. *Org. Biomol. Chem.* **2008**, *6*, 1149−1158.

(52) López-Vallejo, F.; Nefzi, A.; Bender, A.; Owen, J. R.; Nabney, I. T.; Houghten, R. A.; Medina-Franco, J. L. Increased Diversity of Libraries from Libraries: Chemoinformatic Analysis of Bis-Diazacyclic Libraries. *Chem. Biol. Drug Des.* **2011**, *77*, 328−342.

(53) Schreiber, S. L. Target-Oriented and Diversity-Oriented Organic Synthesis in Drug Discovery. *Science* **2000**, *287*, 1964−1969.

(54) Gozalbes, R.; Simon, L.; Froloff, N.; Sartori, E.; Monteils, C.; Baudelle, R. Development and Experimental Validation of a Docking Strategy for the Generation of Kinase-Targeted Libraries. *J. Med. Chem.* **2008**, *51*, 3124−3132.

(55) López-Vallejo, F.; Caulfield, T.; Martínez-Mayorga, K.; Giulianotti, M. A.; Nefzi, A.; Houghten, R. A.; Medina-Franco, J. L. Integrating Virtual Screening and Combinatorial Chemistry for Accelerated Drug Discovery. *Comb. Chem. High Throughput Screening* **2011**, *14*, 475−487.

(56) Ganesan, A. The Impact of Natural Products Upon Modern Drug Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 306−317.

(57) Owens, J.; Lipinski, C. Chris Lipinski Discusses Life and Chemistry after the Rule of Five. *Drug Discovery Today* **2003**, *8*, 12−16.

(58) CRC Dictionary of Natural Products. http://www.crcpress.com (accessed October 2010).

(59) Specs. http://www.specs.net (accessed October 2010).

(60) Clark, R. L.; Johnston, B. F.; Mackay, S. P.; Breslin, C. J.; Robertson, M. N.; Harvey, A. L. The Drug Discovery Portal: A Resource to Enhance Drug Discovery from Academia. *Drug Discovery Today* **2010**, *15*, 679−683.

(61) Chen, X.; Lin, Y. M.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002**, *18*, 130−139.

(62) National Library of Medicine. ChemIDplus Advanced. http://chem.sis.nlm.nih.gov/chemidplus/ (accesed April 2011).

(63) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: A Small-Molecule Screening and Cheminformatics Resource Database. *Nucleic Acids Res.* **2008**, *36*, D351−D359.

(64) Warr, W. A, ChEMBL. An Interview with John Overington, Team Leader, Chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195−198.

(65) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2008**, *36*, D344−D350.

(66) Girke, T.; Cheng, L. C.; Raikhel, N. ChemMine. A Compound Mining Database for Chemical Genomics. *Plant Physiol.* **2005**, *138*, 573−577.

(67) La Chimiothèque Nationale. http://chimiotheque-nationale.enscm.fr/index.php (accessed April 2011).

(68) Del Rio, A.; Barbosa, A. J. M.; Caporuscio, F.; Mangiatordi, G. F. CoCoCo: A Free Suite of Multiconformational Chemical Databases for High-Throughput Virtual Screening Purposes. *Mol. BioSyst.* **2010**, *6*, 2122−2128.

(69) Developmental Therapeutics Program NCI/NIH. http://dtp.nci.nih.gov (accessed April 2011).

(70) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901−D906.

(71) GVK BIO Services/Informatics, Databases GVK BIO. http://www.gvkbio.com/informatics.html (accessed April 2011).

(72) i:lib diverse, inte:ligand. http://www.inteligand.com (accessed April 2011).

(73) Hu, L. G.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins* **2005**, *60*, 333−340.

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

(75) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind Database: Methodologies and Updates. *J. Med. Chem* . **2005**, *48*, 4111−4119.

(76) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.

(77) Chen, X.; Ji, Z. L.; Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2002**, *30*, 412−415.

(78) Zhu, F.; Han, B.; Kumar, P.; Liu, X.; Ma, X.; Wei, X.; Huang, L.; Guo, Y.; Han, L.; Zheng, C.; Chen, Y. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* **2010**, *38*, D787−D791.

(79) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS ONE* **2011**, *6*.

(80) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity, In *Chemoinformatics in Drug Discovery*; Wiley-VCH: New York, 2004; pp 223−239.

(81) Irwin, J. J.; Shoichet, B. K. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(82) Sheridan, R. P. Alternative global goodness metrics and sensitivity analysis: heuristics to check the robustness of conclusions from studies comparing virtual screening methods. *J. Chem. Inf. Model.* **2008**, *48* (2), 426−433.

(83) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 239−255.

(84) Knegtel, R. M.; Wagener, M. Efficacy and selectivity in flexible database docking. *Proteins* **1999**, *37* (3), 334−345.

(85) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *Med. Chem.* **2006**, *49* (23), 6789−6801.

(86) 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* (4), 1504−1519.

(87) Jain, A. N.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133−139.

(88) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.

(89) Sushko, I.; Novotarskyi, S.; Koerner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.

(90) Walker, T.; Grulke, C. M.; Pozefsky, D.; Tropsha, A. Chembench: A Cheminformatics Workbench. *Bioinformatics* **2010**, *26*, 3000−3001.

(91) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49* (3), 678−692.

(92) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488−508.

(93) Bender, A.; Bojanic, D.; Davies, J. W.; Crisman, T. J.; Mikhailov, D.; Scheiber, J.; Jenkins, J. L.; Deng, Z.; Hill, W. A. G.; Popov, M.; Jacoby, E.; Glick, M. Which Aspects of HTS Are Empirically Correlated with Downstream Success? *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 327−337.

(94) Rohrer, S. G.; Baumann, K. Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704−718.

(95) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(96) Pan, Y. P.; Huang, N.; Cho, S.; MacKerell, A. D. Consideration of Molecular Weight During Compound Selection in Virtual Target-Based Database Screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.

(97) Good, A. C.; Oprea, T. I. Optimization of CAMDTechniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.

(98) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(99) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Ann. Rep. in Comput. Chem.*; Elsevier, 2008; Vol. *4*, pp 217−241.

(100) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168−2178.

(101) Doddareddy, M. R.; Klaasse, E. C.; Shagufta; Ijzerman, A. P.; Bender, A. Prospective Validation of a Comprehensive in Silico HERG Model and Its Applications to Commercial Compound and Drug Databases. *ChemMedChem* **2010**, *5*, 716−729.

(102) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.

(103) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(104) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Bin Kang, Z.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding False Positives in Reporter Gene Assays: In Silico Chemogenomics Approaches to Prioritize Cell-Based HTS Data. *J. Chem. Inf. Model.* **2007**, *47*, 1319−1327.

(105) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(106) *SYBYL*; Tripos International: St. Louis, MO, USA.

(107) Pedretti, A.; Villa, L.; Vistoli, G. VEGA: A Versatile Program to Convert, Handle and Visualize Molecular Structure on Windows-Based PCs. *J. Mol. Graphics Modell.* **2002**, *21*, 47−49.

(108) Nobeli, I.; Price, S. L.; Lommerse, J. P. M; Taylor, R. Hydrogen Bonding Properties of Oxygen and Nitrogen Acceptors in Aromatic Heterocycles. *J. Comput. Chem.* **1997**, *18*, 2060−2074.

(109) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

(110) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(111) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation Of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(112) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(113) Hert, J.; Willett, P.; Wilton, D. J. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(114) Dalke Scientific News. http://www.dalkescientific.com/writings/diary/archive/2011/01/20/implementing_cactvs_keys.html (accesed June 2011).

(115) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123−140.

(116) Feund, Y.; Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting, In *Proceedings of the Second European Conference on Computational Learning Theory*; Springer-Verlag, 1995; pp 23−37.

(117) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in Qsar and Qspr. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.

(118) Mattioni, B. E.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting to Generate a Model Ensemble. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 949−963.

(119) Seierstad, M.; Agrafiotis, D. K. A QSAR Model of HERG Binding Using a Large, Diverse, and Internally Consistent Training Set. *Chem. Biol. Drug Des.* **2006**, *67*, 284−296.

(120) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.

(121) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.

(122) Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2d Qsar Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.

(123) Baber, J. C.; William, A. S.; Gao, Y. H.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 277−288.

(124) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(125) Paul, N.; Rognan, D. ConsDock: A New Program for the Consensus Analysis of Protein-Ligand Interactions. *Proteins: Struct. Funct. Genet.* **2002**, *47*, 521−533.

(126) Chen, B.; Mueller, C.; Willett, P. Combination Rules for Group Fusion in Similarity-Based Virtual Screening. *Mol. Inf.* **2010**, *29*, 533−541.

(127) Willett, P. Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143−1152.

(128) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A Distance Geometry Heuristic for Expanding the Range of Geometries Sampled During Conformational Search. *J. Comput. Chem.* **2006**, *27*, 1962−1969.

(129) Vieth, M.; Hirst, J. D.; Brooks, C. L. Do Active Site Conformations of Small Ligands Correspond to Low Free-Energy Solution Structures? *J. Comput.-Aided Mol. Des.* **1998**, *12*, 563−572.

(130) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J. Chem. Inf. Model.* **2006**, *46* (6), 2324−32.

(131) Tawa, G. J.; Baber, J. C.; Humblet, C. Computation of 3D queries for ROCS based virtual screens. *J. Comput.-Aided Mol. Des.* **2009**, *23* (12), 853−868.

(132) 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.

(133) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *J. Chem. Inf. Model.* **2009**, *49* (10), 2303−2311.

(134) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational Analysis of Macrocycles: Finding What Common Search Methods Miss. *J. Chem. Inf. Model.* **2009**, *49*, 2242−2259.

(135) Bostrom, J.; Norrby, P. O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383−396.

(136) Cavasotto, C. N.; Singh, N. Docking and High Throughput Docking: Successes and the Challenge of Protein Flexibility. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 221−234.