Expert Opinion on Drug Discovery

ISSN: 1746-0441 (Print) 1746-045X (Online) Journal homepage: informahealthcare.com/journals/iedc20

How far can virtual screening take us in drug discovery?

Supratik Kar & Kunal Roy

To cite this article: Supratik Kar & Kunal Roy (2013) How far can virtual screening take us in drug discovery?, Expert Opinion on Drug Discovery, 8:3, 245-261, DOI: 10.1517/17460441.2013.761204

To link to this article: https://doi.org/10.1517/17460441.2013.761204

Published online: 21 Jan 2013.

EXPERT OPINION

How far can virtual screening take us in drug discovery?

Supratik Kar & Kunal Roy†
Jadavpur University, Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Kolkata, India

Introduction: Virtual screening (VS) has emerged as an important tool in identifying bioactive compounds through computational means, by employing knowledge about the protein target or known bioactive ligands. VS has appeared as an adaptive response to the massive throughput synthesis and screening paradigm as necessity has forced the computational chemistry community to develop tools that screen against any given target and/or property millions or perhaps billions of molecules in short period of time.

Areas covered: This editorial review attempts to catalog most commonly exercised VS methods, available databases for screening, advantages of VS methods along with pitfalls and technical traps with the aim to make VS as one of the most effective tools in drug discovery process. Finally, several case studies are cited where the VS technology has been applied successfully.

Expert opinion: In recent times, many successful examples have been demonstrated in the field of computer-aided VS with the objective of increasing the probability of finding novel hit and lead compounds in terms of cost-effectiveness and commitment in time and material. Despite the inherent limitations, VS is still the best option now available to explore a large chemical space.

Keywords: computer-aided drug design, docking, high-throughput screening, *in silico*, virtual screening

Expert Opin. Drug Discov. (2013) 8(3):245-261

1. Introduction

In the context of costly clinical trials, animal models and depleted drug discovery pipelines, high-content emerging virtual screening (VS) technologies have appeared as crucial tools in drug discovery research. VS is increasingly used to come up with hits of novel chemical structure from large chemical libraries that yield a unique pharmacological profile. Thus, success of VS is defined in terms of finding interesting new scaffolds rather than many hits. The enthusiasm to embrace rational approaches is triggered in recent years following tremendous advances in the computations and protein crystallography. Thus, VS approaches (**Figure 1**) have gained immense popularity and have become an integral part of the industrial and academic research, directing drug design and discovery [1-5].

2. Emergence of VS and available methods

Computational tools are becoming increasingly important to integrate structural data with the more traditional lead optimization techniques. Among them, VS is bringing a more cost-effective and sensible approach to drug discovery. An effort is made in this article to present the commonly available VS methodologies and to categorize them into various groups [6-10]. The theory and principles behind these VS procedures are thoroughly discussed in **Table 1**.

informa
healthcare

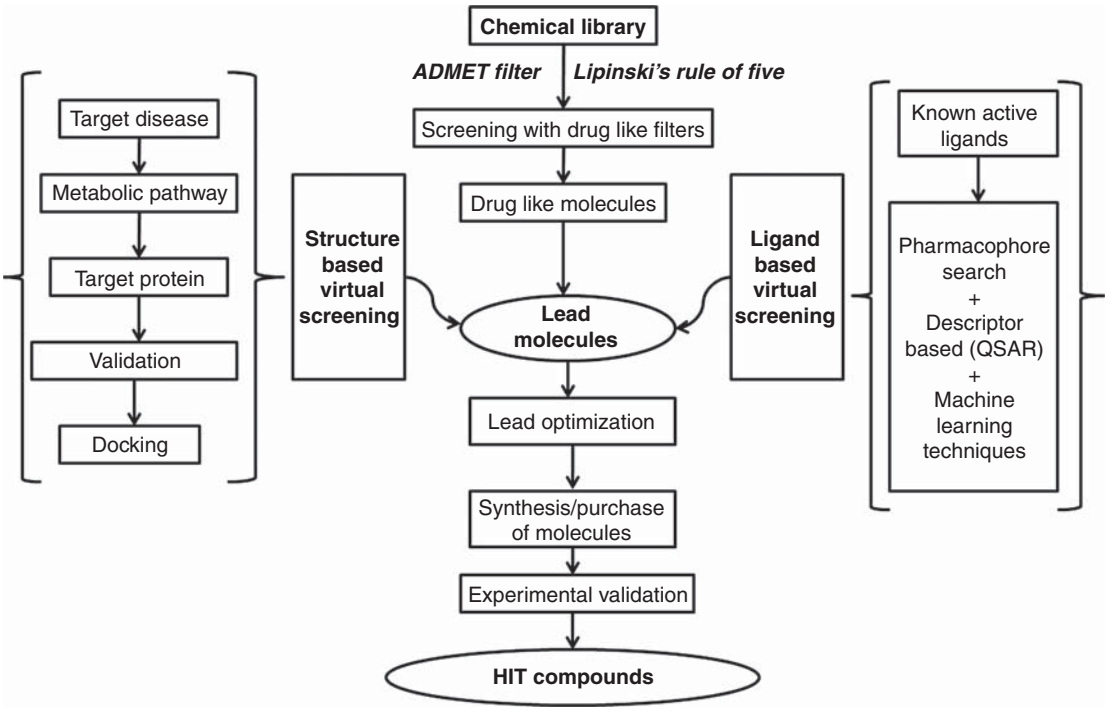

Figure 1. Schematic representation of commonly practiced VS process.

3. Necessity/advantages of VS

The use of complementary experimental and VS techniques increases the chance of success in many stages of the discovery process. VS has emerged as a reliable, cost-effective and time-saving technique for the discovery of lead compounds in recent times [11,12]. The main advantages of this method compared to laboratory experiments are noted below.

3.1 Cost-effective

As no compounds have to be purchased externally or synthesized by a chemist, VS is one of the most cost-effective methods in the preliminary stage of the drug discovery processes.

3.2 Time effectual

Time required for synthesis can be extremely hefty, especially in case of large database with millions of chemical compounds. But using computational tools, VS approach is always a time effectual tool in drug development.

3.3 Labor efficient

Synthesis and bioassays are always laborious and chance of getting false positives is always present after spending lots of physical and mental labor. Though one cannot deny the chance of getting false positives in case of VS, it is always a labor efficient tool in the drug development process.

3.4 Sensible alternative

It is possible to investigate compounds that have not been synthesized yet. Conducting high-throughput screening (HTS) experiments is expensive, time-consuming and laborious for huge number of chemicals. As a result, VS is always a sensible option to reduce the initial number of compounds before using HTS methods.

4. Pitfalls, technical traps and cautionary notes of VS

Identification of the target protein and the active site with an ideal ligand is not sufficient to reach any logical end in a drug discovery process. There are many obstructions, which come in the way of designing a new drug compound to a final drug molecule. Therefore, it is indispensable to take account of small and subtle factors which play a crucial role in making a drug candidate active. Therefore, exhaustive computational treatments still fall short of providing a reliable answer [13-15]. We classified the commonly occurred drawbacks into the following six categories.

4.1 Erroneous assumptions and expectations

4.1.1 Predicting the wrong binding pose

Docking-based VS can accidentally produce the right result for the wrong reasons; that is, it can correctly assign high scores to many true hits but still predict the wrong binding poses.

Table 1. Commonly practiced VS methods.

Methods	Types	Sub-types	Classification	Remarks
Structure-based VS	Docking	Based on searching methods	Monte Carlo (MC)	One of the most widely used simulated annealing procedures which can generate states of low energy conformations. The system makes random moves and accepts or rejects each conformation based on Boltzmann probability. The random motion of the ligand allows for exploration of the local search space, and the decreasing temperature of the system acts to drive it to a minimum energy
			Genetic algorithm (GA)	An adaptive heuristic search technique based on the evolutionary ideas of natural selection and genetics. In a GA, there is a population of solutions that undergo mutation and crossover transformations. The newly generated solutions undergo selection, biased toward the fit among them. The algorithm maintains a selective pressure toward an optimal solution, with a randomized information exchange permitting exploration of the search space. A range of programs implements GA for docking, including GOLD, AutoDock and DARWIN
		Based on scoring functions	Force-field-based	It is generally accurate in estimating binding free energies, when occupied with free energy perturbations or thermodynamic integration methods. Force field-based scoring functions sum-up the contributions from different interaction energy categories such as hydrogen bond, electrostatic and van der Waals between ligand and protein atoms. It is crucial to assess the relative binding strength of screened compounds that represent the ligand conformations and hence the potential energy functions (force fields) used during the docking process is necessarily simplified
			Empirical	The design of empirical scoring functions is based on the idea that binding energies can be approximated by a sum of individual uncorrelated terms. These scoring functions are fit to reproduce experimental data, such as binding energies and conformations, as a sum of several parameterized functions. The appeal of empirical functions is that their terms are often simple to evaluate, but they are based on approximations similar to force-field functions
			Consensus	Due to the imperfections of current scoring functions, a recent trend has been the introduction of consensus scoring schemes and it involves highly disparate properties in order to improve performance in VS. Consensus scoring combines information from different scores to balance errors in single scores and improve the probability of identifying 'true' ligands. An exemplary implementation of consensus scoring is X-CSCORE, which combines GOLD-like, DOCK-like, ChemScore, PMF and FlexX scoring functions
			Knowledge-based	Knowledge-based scoring functions are designed to reproduce experimental structures rather than binding energies. Protein–ligand complexes are modeled using relatively simple atomic interaction–pair potentials. A number of atom–type interactions are defined depending on their molecular environment. So, in common with empirical methods, knowledge-based scoring functions attempt to implicitly capture binding effects that are difficult to model explicitly. A major advantage is their computational simplicity, which permits efficient screening of large compound databases. A disadvantage is that their derivation is essentially based on information implicitly encoded in limited sets of protein–ligand complex structures

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
		Flexibility of ligand and target	-	During docking process, almost all the software packages allow the flexibility of the ligand, where as they assume that the protein target is held fixed in its crystal structure conformation. Software programs, such as Discovery studio (Accelrys), Omega, Confort (Openeye Scientific Software), Checkmol, rapidly generate large conformational database, and each conformer is rigidly docked into the target-binding site. The extent of flexibility is dependent on the number of rotatable bonds present in the ligand. In FlexXE and Affinity programs, flexibility is given to both ligand and the receptor site. The other important factors in structure-based drug design are determination of conformational changes in receptor during ligand binding, protonation states and the tautomer enriched databases. An algorithm called the ICM-flexible receptor-docking algorithm (IFREDA) is used in estimating protein flexibility in VS
		Based on programs	DOCK	It systematically describes the geometries of ligands and binding sites by sets of spheres, attempting to fit each compound from a database into the binding site and the spheres could be overlapped by means of an approximate clique-detection procedure. Steric matching-scores with electrostatic and molecular mechanics interaction energies are considered for the ligand–receptor complex
			AutoDock	A script-driven flexible automated and random search docking technique operated by altering the ligand or a subset of ligand with several rotatable bonds to predict the binding interaction between small molecules to the known receptor 3D structure. The robustness of AutoDock can be attributed to the MC simulated annealing, evolutionary, genetic and Lamarckian GA methods
			GOLD	It uses a flexible docking mode for small molecules into protein binding site which utilizes GA for the conformational search that forms a powerful tool for screening and identification of novel lead compounds. It is very highly regarded within the molecular modeling community for its accuracy and reliability and its GA parameters are optimized for wide range of VS applications
			Darwin	DARWIN uses the combination of a GA with a gradient minimization search strategy
			FlexX	FlexX uses a pose-clustering algorithm to classify the docked ligand conformers, where the placement of rigid core fragment is based on interaction geometry between fragment and receptor groups. Prior to docking, FlexX cuts the ligand at rotatable bonds into pieces, places a base fragment into the active site and incrementally builds up the ligand again, using the other pieces. For a protein with known 3D structure and a small ligand molecule, FlexX predicts the geometry of the protein–ligand complex and estimates the binding affinity
			Grid-based ligand docking with energetics (GLIDE)	It approximates a close and complete systematic search for the conformational, orientational and positional space of the docked ligand. GLIDE uses a series of hierarchical filters for searching possible locations of the ligand in the active-site region of the receptor. A grid representation of the shape and properties of the receptor is used that progressively scores for ligand posing

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
			Screening for ligands by induced-fit docking (SLIDE)	It represents a general approach to organic and peptidyl database screening. It can handle large binding-site templates and uses multistage indexing to identify feasible subsets of template points for ligand docking. An optimization approach based on mean-field theory is applied to model induced-fit complementarities, balancing flexibility between the ligand and the protein side chains
			FRED	FRED's docking strategy is to exhaustively score all possible positions of each ligand in the active site. The exhaustive search is based on rigid rotations and translations of each conformer. It takes a multi-conformer library or database of one or more ligands, a target protein structure, a box defining the active site of the protein and several optional parameters. Various options are available for optimization with respect to the built-in scoring functions: optimization of hydroxyl group rotamers, rigid body optimization, torsion optimization and reduction of the number of poses that are passed on to the next scoring function. The ligand conformers and protein structure are treated as rigid during the docking process
			Hammerhead	Hammerhead is suitable for screening large databases of flexible molecules by binding to a protein of known structure. The approach is completely automated from the elucidation of protein binding sites, through the docking of molecules, to the final selection of compounds
Ligand-based VS	Small molecule alignment	-	Genetic Algorithm Similarity Program (GASP)	The reference template is managed to be flexible and the test molecule is flexible. GASP is a GA-based program where few most active molecules are taken as reference molecules and automatically it takes the common features of these reference ligand molecules to compare with the test molecule. This is useful to study cases where conformational preferences cannot be assigned to molecules. It tracks the information about the features of molecule-like mapping between hydrogen-bond donor, acceptor, lone pair and ring center features in pairs of molecules and angles of rotation
			Molecular electrostatic potential (MEP)	It can be used to assess the molecular similarity in ligand-based drug designing. Molecular fields of the test and reference molecule are aligned in such a manner to maximize their similarity and this is achieved by application of GA. It provides flexibility in both test and reference molecules and aligns them in manner so as to maximize their overlap
			Flash-Flood	It is another fragment-based similarity search and alignment method for flexible molecules, where user-defined properties are used for comparison of the fragment pair. In Flash-Flood, the field properties are characterized using descriptor and the conformational feature generated alignment with each other

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
	Descriptor-based screening	1D- and 2D-descriptors	Binary descriptor	In binary descriptor representation, the presence of structural properties for each position of lead molecule is narrated by means of a Boolean bit set to 'one' otherwise to 'zero'. It may indicate different qualitative properties, such as presence or absence of specific functional groups or specific bonds (such as of hydrogen bonds or sulfide bonds etc.) or it may indicate the number of specific bonds or functional groups present i) *Structural keys*: Structural keys are array of Boolean values, which give information about presence or absence of a 2D fragment and are generated using fragment dictionary. MACCS keys are the structural keys, where it is possible to take into account the frequency of occurrence of each fragment instead of simple presence or absence. So it is better to use MACCS keys for fast screening ii) *Molecular fingerprints*: The structural keys have a limitation that it requires sufficient set of predefined fragments. In order to avoid this inconvenience, molecular fingerprints are introduced. These descriptor models for a molecule are generated from the molecule itself and each bit represents a sequence of linked atoms with specified atom types, called a path. The fingerprinting algorithm generates a list of all paths up to a specified length. In molecular fingerprint a particular bit represents not a single specific bond path, but several paths. Examples are Daylight fingerprint and UNITY fingerprint
			Real-value descriptor	The real value descriptor vectors represent the pharmacophoric site of a lead compound by generating a hologram. For estimating similarity between two different molecules different functions, such as Tanimoto coefficient and cosine coefficient, are used. A number of distance functions have been proposed, including the well-known Manhattan and Euclidean metrics for real-valued descriptors and the Tanimoto coefficient for binary descriptors. For example, the VolSurf 112 program represents the 3D molecular field interactions between the molecules using 2D descriptor Mapping to activity class-specific descriptor value ranges (MAD) systematically compares descriptor value range of reference compound activity classes with those of test database compounds in order to identify descriptors having activity class-specific settings. Moreover, MAD utilizes original descriptor values and not binary transformed or otherwise simplified descriptor representations. Improving it further, dynamic MAD (DynaMAD) is introduced which include dynamic descriptor selection. In MAD, descriptor selection was carried out during the initial step, leading to a constant descriptor number for mapping. By contrast, in DynaMAD, the number of selected descriptors increase in a stepwise manner and alternate with compound mapping

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
			Feature tree	Feature trees are found to be more descriptive in comparison to linear descriptor as they also represent the relative position of functional group on molecular surface. By analyzing feature tree, one can estimate the physicochemical properties of the molecule. The similarity between two feature trees can be determined by comparing them with respect to their nodes. Even though feature tree comparison gives more accurate result in comparison to linear descriptors, this method is very complex and time-consuming. For analysis of large databases, the feature tree descriptors are now converted to linear descriptors. NIPALSTREE, a recently developed hierarchical clustering algorithm carries out large database analysis in high-dimensional space
		3D descriptors	-	The estimation of similarity in descriptor-based analysis is also based on different framework of 3D descriptors and the different coefficients used in this search procedure. The 3D descriptors can be generated using different programs, such as 3D-FEATURE based on different hydrophobic groups, hydrogen bond acceptor and hydrogen bond donor. The ligands can also be transformed from 2D to 3D by means of programs such as CORINA. Numerous descriptors can also be generated by taking into consideration the functional groups and primary shape properties. One program that successfully uses the 3D descriptor is distributed information search component (DISCO). In this program, the pharmacophoric points are identified on the lead and different bond distance are taken as a parameter to generate descriptors Comparative molecular field analysis (CoMFA) is a 3D QSAR technique based on data from known active molecules. The aim of CoMFA is to derive a correlation between the biological activity of a set of molecules and their 3D shapes, electrostatic and hydrogen bonding characteristics. This correlation is derived from a series of superimposed conformations, one for each molecule in the set. These conformations are presumed to be the biologically active structures, overlaid in their common binding mode. Each conformation is taken in turn, and the molecular fields around it are calculated. The fields, usually electrostatic and steric (van der Waals interactions) are measured at the lattice points of a regular Cartesian 3D grid; the lattice spacing is typically 2 Å. The 'measured' interaction is between the molecule and a probe atom (a sp3-hybridized carbon with +1 charge) Comparative molecular similarity indices analysis (CoMSIA) method was developed to improve limitations of steric and electrostatic fields in CoMFA. CoMSIA can calculate the hydrophobic descriptor as well as steric and electrostatic descriptors. In CoMSIA, the probe atom having a radius of 1 Å, charge of +1 and hydrophobicity of +1 is placed at the intersections of a surrounding lattice. The steric, electrostatic and hydrophobic similarities of each molecule with the probe atom are calculated using Gaussian-type functions. Hydrogen bonding properties are also studied using CoMSIA

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
		Scaffold hopping	Molprint technique	Scaffold hopping is a recently developed advanced similarity searching procedure. A set of surface points is defined in terms of binary descriptors indicating the presence or absence of particular feature at a specific position. Next, the descriptor information is processed further, by information-gain based feature selection, which in turn depends on the entropy of total subsets that are partitioned with respect to the feature under consideration. Finally, the selected compounds are given scoring using a Naive Bayesian Classifier
			Feature point pharmacophores (FEPOPS)	It is a fully automated scaffold method to simplify flexible 3D chemical descriptors. It is a combination of field-based similarity and pharmacophore-based similarity searching techniques
			Chemically advanced template search (CATS)	This scaffold hopping program software can be used to perform similarity searching in a collection of small molecules. It represents lead molecules in terms of 2D topological pharmacophore descriptors and compares with the reference molecules
	Pharmacophore similarity search	-	P. Goodford's grid program	It is based on location-specific properties of molecules. Pharmacophoric search also sometimes carried out based on point comparison method, where in the alignment is done by using 3D space. An exhaustive grid-based search is done by X-AUTOFIT program
			BRUTUS	Another fast grid-based algorithm namely BRUTUS carries out rigid body molecular superposition and does the alignment, taking field information derived from charge distribution. This can also identify structurally diverse compounds from large database and can be operated in normal desktop computer
			FBSS	In some similarity search programs, such as FBSS, the bioactivity is assessed by taking into consideration the atoms and bond features of lead molecules in restricted volume rather taking the whole volume of ligand to consideration
	Recursive partitioning (RP)	-	CerBeruS	This iterative screening process based on Daylight fingerprint screens most similar compounds by structural similarity features. In successive steps of iterative screening all lead clusters are selected for retesting. Although the accuracy rate is higher in this screening, this technique is only for medium-sized molecules
			Multi-computer automated evaluation (MCASE)	It uses RP technique for screening. In this technique also, different statistical variables are used in different steps to screen the database. MCASE carry out binary activity classification by correlating different features
			PGLT	RP principle is used for data mining. PGLT is useful for molecular screening of ligand when enough information regarding them is available
	Graph-based similarity assessment	-	Rapid similarity calculation	It is based on similarity search with iterative graph matching procedure. Similarity search detects the level of similarity and regulates the selection of lead compound, based on user defined similarity level and the graph match. The resultant graphs are compared by using several similarity coefficients
			Reduced graph (RG)	It provides information about topological pharmacophore. RG considers the molecular subgroups for comparison which takes part in ligand–receptor protein interaction. The similarity search here is carried out by different advance graph

Table 1. Commonly practiced VS methods (continued).

Methods	Types	Sub-types	Classification	Remarks
				matching techniques. The RG matching process is also considered as one of the scaffold hopping process. The RG is encoded using binary fingerprint. RG can also represent the ligand–receptor interaction by using whole molecular ligand descriptor. Its performance is quite reliable when compared with the other VS techniques, such as Daylight fingerprints, FTrees, UNITY and various FlexX docking protocols
	Machine learning techniques	Support vector machine (SVM) technique	-	SVM predicts the bioactivity of compound from its conformational feature. It predicts the bioactivity by representing the lead in *n*-dimensional real space using molecular descriptor and fingerprint technology where '*n*' represents number of features or attributes. SVM approach is based on fuzzy logic fingerprint
		Binary Kernal discrimination (BKD)	-	BKD is a recently developed computational approach. In BKD, the molecule is represented as 2D fragment bit-string. It consists of three components. First, structural representation section, second similarity searching section using different coefficients and third section with different weighting schemes for lead compounds
		Self-organizing maps (SOM)	-	SOM or Kohonen neural network is one of the basic types of artificial neural networks. Its architecture represents a 2D grid of connected neurons, which are multidimensional vectors. The projection or learning of network runs in two steps, the first step is the selection of the winning neuron and the second step is the self-organization of the map. The mapping is topology preserving which means that similar objects in descriptor space are located close to each other (or even on the same neuron) but it is not metric preserving. A map is not only a picture of original space but also a model
		Multilayer perceptron neural networks (MLPs)	-	MLP is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes a supervised technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that is not linearly separable
		Counter-propagation neural network (CPNN)	-	The CPNN implements input and output variables differently. The input layer has the same structure as in SOM and the output layer is situated beneath. The difference to SOM lies in the learning strategy. The prediction runs over two steps. In the first step, the object is located into input layer on the neuron with the most similar weights. In the second step, the position of that neuron is projected to the output layer, which gives the predicted output value

4.1.2 Variable water-mediated binding interactions
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.

4.1.3 Single versus multiple/allosteric binding pockets
Both structure- and ligand-based VS approaches have the intrinsic shortcoming that they cannot identify bioactive ligands for binding pockets which are not explicitly docked against or implicitly represented in the training set.

4.1.4 Post-VS compound selection
A better approach might be to cluster the compounds into related families and select a small number of molecules from each cluster. While this approach is perfectly sensible, it introduces a bias in the VS protocol, rendering comparisons difficult.

4.1.5 Prospective validation
External validation on new data sets is not considered during the model development and it is rarely discussed in VS.

4.1.6 Drug-likeness
Many VS approaches are based on 'drug-like' compounds, as defined by Lipinski in his 'rule-of-five' work. 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.

4.1.7 Chemical characteristics
There are at least three areas of context-dependent chemical characteristics that are relevant to VS: i) tautomeric form, where the selection of a wrong tautomer can misguide the assignment of hydrogen bond acceptors (HBA) or donors, ii) ionization, where the protonation state of chemical groups at physiologically relevant pH can be miscalculated and iii) chirality, where for racemic structures one needs to calculate the conformations for all possible chiral configurations.

4.2 Regarding data design and content
4.2.1 Hit rate in standard data sets
Two factors that complicate benchmarking of VS algorithms are the size and diversity of the chemical libraries. Standard libraries are either too small or they contain too many closely related analogs or often both.

4.2.2 'Bad or problematic' molecules
Datasets may include molecules that contain chemically reactive groups or other undesirable functionalities that interfere with the HTS detection techniques. 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.

4.2.3 Feature weights
Ligand-based VS, based on a single query, typically shows equal importance on all parts of the molecule. However, some substructural features may not be required for activities against a specific target of interest.

4.3 Concerning conformational sampling as well as ligand and target flexibility
4.3.1 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.

4.3.2 Size of conformational ensemble
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.

4.3.3 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. The number of conformations accessible to a molecule largely depends on its size and flexibility.

4.3.4 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.

4.3.5 Target flexibility
Indeed, it is not only the ligands but also the biological targets that are flexible. Protein flexibility is probably the most unexploited aspect of VS.

4.3.6 Assumption of ligand overlap
In 3D shape-based VS, most programs attempt to maximize the overlap between the query and the database molecules. 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.

4.4 Choice of software
4.4.1 Input–output errors and format incompatibilities
An ordinary but serious problem is error introduced while interconverting different molecular formats. 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.

Table 2. List of successful case studies over the years (representative list).

Target/molecule	Tools/software	Methods/schemes	Link
5HT2c receptor	FlexX, FRED, MOE, FTress, Daylight fingerprints	Incremental docking, ScreenScore, PMF, FlexX for scoring	http://onlinelibrary.wiley.com/doi/10.1002/prot.20651/pdf
Acetohydroxyacid synthase (AHAS)	DOCK 4.0, AutoDock3.0, XLOGP, SYBYL6.9, GRID SPHGEN	Docking, VS based on the binding site	http://dx.doi.org/10.1016/j.bmcl.2006.06.057
Acetyl cholinesterase (AChE) inhibitors	Ligand Scout	Structure-based pharmacophore model. A 3D multi-conformational database with more than 110,000 natural products	http://pubs.acs.org/doi/abs/10.1021/jm049655r
Acetyl cholinesterase	CAL GRID, AMBER, EVE-MAKE, ADAM & EVE	ADAM & EVE method takes into consideration the flexibility of molecules by fully exploring conformation space. EVE-MAKE is used for flexible docking prior to 3D screening. AMBER is used for calculating force field energy	http://pubs.acs.org/doi/abs/10.1021/jm030605g
Aldo-Keto reductases (AKR1C1 and AKR1C3)	FlexX 3.1, Gaussian 09	Docking and inhibition assay	http://pubs.acs.org/doi/abs/10.1021/jm300841n
Alpha 1A receptor	GOLD, Catalyst	Homology modeling, GA. application of 2D filters and pharmacophore	http://pubs.acs.org/doi/abs/10.1021/jm0491804
Androgen receptor activation function-2 (AF2)	GLIDE SP, MOE	Docking, consensus scoring and bioassay	http://pubs.acs.org/doi/abs/10.1021/jm200532b
Angiotensin converting enzyme-2	Catalyst, eHiTS	Pharmacophore model and flexible ligand docking	http://pubs.acs.org/doi/abs/10.1021/ci0503614
Antagonists for the G protein-coupled NK3 Receptor	SYBYL 8.1, ROCS 3.0	Sequential similarity analysis followed by CoMFA	http://pubs.acs.org/doi/abs/10.1021/jm1010012
Aspartic protease rennin	LigandFit/Cerius, LigScore1, LigScore2, PLP1, PLP2, JAIN, PMF, LUDI	Shape-based VS and consensus scoring	http://pubs.acs.org/doi/abs/10.1021/ci0342728
B-secretase (BACE1)	FlexX, FlexX-Pharm	Incremental build scoring functions are applied	http://pubs.acs.org/doi/abs/10.1021/jm0491804
Beta-catenin	QXP docking, NMR and ITC filtering	Combination of docking and biophysical screening	http://onlinelibrary.wiley.com/doi/10.1002/prot.20955/pdf
Biogenic amine-binding GPCRs	GOLD, FlexX-Pharm, FTREE, Catalyst	GA, incremental docking and homology modeling	http://pubs.acs.org/doi/abs/10.1021/jm049133b

Table 2. List of successful case studies over the years (representative list) (continued).

Target/molecule	Tools/software	Methods/schemes	Link
Cannabinoid receptors	SYBYL, GOLD	GA-based screening	http://pubs.acs.org/doi/abs/10.1021/jm060394q
CB2 receptor	PROCHECK, GOLD and SYBYL CSscore	Structure-based screening	http://pubs.acs.org/doi/abs/10.1021/jm050565b
Checkpoint kinase-1 (Chk-1 kinase)	Leatherface Daylight tool kit, Corina, Omega, FlexX-Pharm	Incremental docking and generation of 3D pharmacophore. CScore module and consensus scoring	http://pubs.acs.org/doi/abs/10.1021/jm030504i
COX-2	CoMFA, SOM, LVQ, Binary, Decision Tree, PLS, BRANN, GA-KNN, Linear, MOE	Structure-based and ligand-based VS. Scoring based on 2D QSAR consensus prediction. Compounds from the NCI molecular database	http://pubs.acs.org/doi/abs/10.1021/ci0341565
Cyclin-dependent kinase 2(CDK2)	GOLD, QXP	GA and QXP docking. 10 different scoring functions are used	http://onlinelibrary.wiley.com/doi/10.1002/prot.20473/pdf
Cysteine protease	Gold, SYBYL, Cerius2,Catalyst, Concord	GA- and pharmacophore-based screening	http://pubs.acs.org/doi/abs/10.1021/jm0505765
Dengue viral NS2B-NS3 protease	DOCK 4.0, GOLD 3.0	Small molecule-based scaffold hopping and structure-based VS	http://pubs.acs.org/doi/abs/10.1021/jm300146f
Dihydrofolate reductase (DHFR)	Hydropathic interactions (HINT), GRID	Computational titration analysis in order to optimize the ionization states of residues	http://dx.doi.org/10.1016/j.bmc.2006.09.050
Dipeptidyl peptidase IV	GLIDE	Pharmacophore-based screening and flexible ligand docking	http://pubs.acs.org/doi/abs/10.1021/jm0505866
Epidermal growth factor receptor kinase (EGFR)	Gold, Ligandfit, 128 EGFR kinase inhibitors	GA, MC simulations, consensus scoring	http://pubs.acs.org/doi/abs/10.1021/ci049676u
ER_{β}	Catalyst GOLD, CCDC	Genetic optimization, Gold score	http://pubs.acs.org/doi/abs/10.1021/jm0490538
ERG2 and EBP	Catalyst, Sybyl	Pharmacophore-based screening	http://pubs.acs.org/doi/abs/10.1021/jm049073%2B
Erythropoietin-producing hepatocellular B2 (EphB2) receptor	GOLD, DISCO. SYBYL	GA, Pharmacophore generation and CSscore	http://pubs.acs.org/doi/abs/10.1021/jm0492204
Estrogen receptors	GEMDOCK	Flexible molecular docking and pharmacophore-based scoring function	http://onlinelibrary.wiley.com/doi/10.1002/prot.20387/pdf
Falcipain-2, Falcipain-3	GOLD	Three consecutive docking methods on ChemBridge database	http://pubs.acs.org/doi/abs/10.1021/jm0493717
Glycogen synthase kinase-3a (GSK-3a)	FlexX, FlexX-Pharm, FlexE, Sybyl	Structure-based VS	http://pubs.acs.org/doi/abs/10.1021/jm050504d
Growth hormone secretagogue agonist	Fingerprints and cell-based partitioning	Similarity searching and cell-based compound partitioning	http://pubs.acs.org/doi/abs/10.1021/jm040103i
HIV protease	Omega, MOE	Pharmacophore mapping.	http://pubs.acs.org/doi/abs/10.1021/ci050511a

Table 2. List of successful case studies over the years (representative list) (continued).

Target/molecule	Tools/software	Methods/schemes	Link
Human aldose reductase (HAR)	FlexX, Corina	Hierarchical screening using incremental docking and generation of pharmacophore DrugScore	http://onlinelibrary.wiley.com/doi/10.1002/prot.20057/pdf
Human AICAR transformylase	AutoDock	Structure-based screening of NCI database	http://pubs.acs.org/doi/abs/10.1021/jm049504o
Human factor Xa	Ligandscout, Catalyst, iLib Diverse.	Ligand-based, structure-based screening and generation of virtual combinatorial library	http://pubs.acs.org/doi/abs/10.1021/ci049778k
Human murine double minute 2(MDM2)-p53	GOLD, ChemScore	GA	http://pubs.acs.org/doi/abs/10.1021/jm060023%2B
Human platelet type 12-and reticulocyte 15-lipoxgenase-1	GLIDE, ChemScore	Flexible ligand docking	http://pubs.acs.org/doi/abs/10.1021/jm050639j
Kinesin spindle protein (KSP)	HypoGen algorithm within catalyst	3D pharmacophore model development	http://dx.doi.org/10.1016/j.bmcl.2006.08.061
Kv1.5 blockers	CATS, GRID, SYBYL 6.8, CORINA, VolSurf3.0	Ligand-based screening	http://pubs.acs.org/doi/abs/10.1021/ci0400011
Melanin-concentrating hormone 1 receptor (MCH-1R)	Daylight SMARTS and Fingerprints, Unity fingerprints and FlexS	2D, 3D similarity and substructure analysis	http://pubs.acs.org/doi/abs/10.1021/jm040762v
Metabotropic glutamate receptor subtype 4 (mGlu4R)	Ligand and DREDING, DOCK score	Receiver operating characteristic (ROC) curve method	http://pubs.acs.org/doi/abs/10.1021/jm049092j
Metalloenzymes	DOCK 3.5.54, DISTMAP, CHEMGRID, AMBER, Delphi	Incremental, noncovalent scoring function. MDDR and ACD databases were used	http://pubs.acs.org/doi/abs/10.1021/bi050801k
Neurokinin-1 (Nk-1) (GPCRs)	MOBILE, FlexX-Pharm, and MAB force field	Structure-based VS and receptor model generation	http://pubs.acs.org/doi/abs/10.1021/jm0311487
Non-peptide malignant brain tumor (MBT) antagonists	Discovery Studio 2.5	3D pharmacophores and structure-based screening	http://pubs.acs.org/doi/abs/10.1021/jm1007374
Peroxisome proliferator-activated receptor	Catalyst, GOLD	Shape similarity searching. Maybridge database was used	http://pubs.acs.org/doi/abs/10.1021/jm051129s
P. falciparum dihydroorotate dehydrogenase (PfDHODH)	Discovery Studio 2.1, Vlife MDS 4.1	3D pharmacophore, docking and SCOPE model used for VS	http://onlinelibrary.wiley.com/doi/10.1002/minf.201200045/pdf
PH domain leucine-rich repeat protein phosphatase (PHLPP)	GLIDE, MODELER	Homology modeling, docking, cell culture and immunoblotting	http://pubs.acs.org/doi/abs/10.1021/jm100331d
Protein arginine methyltransferases (PRMTs)	Accelrys Discovery Studio 2.1	Pharmacophore-based VS. SPECS database from the ZINC database	http://pubs.acs.org/doi/abs/10.1021/jm300521m
Protein kinases	p-SIFt, FlexX		http://pubs.acs.org/doi/abs/10.1021/jm049312t

Table 2. List of successful case studies over the years (representative list) (continued).

Target/molecule	Tools/software	Methods/schemes	Link
		Profile-based scoring, interaction-based analysis, incremental docking	
Protein phosphatase 2C	AutoDock	Grid-based screening	http://pubs.acs.org/doi/abs/10.1021/jm051033y
SARS 3C-like proteinase	DOCK 4.01,LigBuilder	Incremental build, pharmacophore generation and consensuous scoring. NCI, MDDR, ACD databases were used	http://pubs.acs.org/doi/abs/10.1021/jm050090o
Tat-TAR RNA	SQUID, CATS 3D, MOE	Ligand-based VS. Alignment-free pharmacophores method and fuzzy pharmacophores approach	http://onlinelibrary.wiley.com/doi/10.1002/cbic.200400376/pdf
Thymidine monophosphate kinase (TMP)	Cerius, GOLD, catalyst	GA, combination of docking and pharmacophore model	http://pubs.acs.org/doi/abs/10.1021/ci050064z
Topoisomerase I inhibitor	GOLD, AutoDock, MVD	Structure-based screening. Specs database was used	http://pubs.acs.org/doi/abs/10.1021/jm100387d
Topoisomerase II Alpha	Catalyst, Cerius 2	Pharmacophore generation and screening based on NCI database	http://pubs.acs.org/doi/abs/10.1021/jm049745w
Trypanothione reductase	FlexX and ProPose, MOLOC	Structure-based screening	http://pubs.acs.org/doi/abs/10.1021/jm050027z
Tyrosinase inhibition	CARDD module in TOMOCOMD	Screening using non-stochastic and stochastic bond-based linear indices	http://pubs.acs.org/doi/abs/10.1021/jm060526f

Table 3. Listing of available online compound databases for screening (representative list).

Compound database	Availability	No. of compounds	Website
ACD	Commercial	3,870,000	http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html
Asinex	Commercial	550,000	http://www.asinex.com
Binding database	Public	284,206 small ligands with 648,915 binding data, for 5,662 protein targets	http://www.bindingdb.org
Chem ID	Public	388,000	http://chem.sis.nlm.nih.gov/chemidplus/
ChemBank	Public	800,000	http://chembank.broadinstitute.org
ChEMBL db	Public	658,075 differing bioactive compounds and 8,091 targets	https://www.ebi.ac.uk/chembldb/
ChemBridge	Commercial	700,000	http://www.chembridge.com
ChemDiv	Commercial	1,500,000	http://www.chemdiv.com
Chemical entities of biological interest (ChEBI)	Public	584,456	http://www.ebi.ac.uk/chebi/init.do
ChemMine	Public	6,200,000	http://bioweb.ucr.edu/ChemMineV2/
ChemNavigator	Commercial	55,300,000	http://www.chemnavigator.com
ChemSpider	Public	26,000,000	http://www.chemspider.com
Chimiotheque nationale	Public	44,817 compounds	http://chimiotheque-nationale.enscm.fr/index.php
CoCoCo	Public	6,957,134 molecules and more than 14,514 conformations	http://cococo.unimore.it/tiki-index.php
Developmental therapeutics program (DTP)	Public	4,73,965	http://dtp.nci.nih.gov/
DrugBank	Public	6,827 drugs, 4,477 nonredundant protein sequences	http://www.drugbank.ca
Enamine	Commercial	1,700,000	http://www.enamine.net
FDA database	Public	Drugs@FDA includes most of the drug products approved since 1939	http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
GLIDA	Public	G-protein-coupled receptors (GPCRs) related chemical genomics database	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/index.php
GVK BIO	Commercial	Focused libraries with target inhibitor	http://www.gvkbio.com/informatics.html
iLib Diverse	Commercial	Drug-like fragment set for combinatorial library generation	http://www.inteligand.com/
Interbioscreen	Public	440,000 synthetic and 47,000 natural	http://www.ibscreen.com/index.htm
Maybridge	Commercial	56,000	http://www.maybridge.com
MDDR	Commercial	150,000	http://accelrys.com/products/databases/bioactivity/mddr.html
Mother of all databases (MOAD)	Public	14,720 ligand–protein complexes, 4,782 structures with binding data, 7,064 ligands	http://www.bindingmoad.org
NCI	Public	140,000 million	http://dtp.nci.nih.gov/index.html
PDB bind	Commercial	3,214 ligand–protein complexes	http://www.pdbbind.org/
PubChem	Public	49,875,000	http://pubchem.ncbi.nlm.nih.gov
Specs	Commercial	240,000	http://www.specs.net
Super drug database (SDD)	Public	2,396 compounds with 1,08,198 conformers	http://bioinf.charite.de/superdrug/
TCM	Public	32,000	http://tcm.cmu.edu.tw
Therapeutic target database	Commercial	1,906 targets, 5,124 drugs	http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp
WOMBAT	Commercial	305,727 molecule, 1,966 targets	http://www.sunsetmolecular.com
ZINC	Public	13,000,000	http://zinc.docking.org

4.4.2 *Molecule preparation*

Adding implicit hydrogen atoms and assigning the correct charges can easily be forgotten in many VS algorithms that critically depend on these parameters.

4.4.3 *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 HBA simultaneously.

4.4.4 *Fingerprint selection and algorithmic implementation*

In similarity-based screening, performance depends critically on the choice of descriptors. Also, as with all software, descriptors may be implemented in a different way within different software packages.

4.5 Selection of suitable database library for VS

Chemically diverse libraries are particularly attractive for identifying novel scaffolds for new or relatively unexplored targets, such as those from diversity-oriented synthesis. If the goal of the screening is directed at a specific target family, one may use target-oriented synthesis, focused or targeted libraries.

4.6 Single predictors versus ensembles

A common experience for the VS practitioners is that different screen methods retrieve different molecules from the same database. Thus, one needs to run several VS methods for any given target and additively collect the outcome.

5. Case studies: successful application of VS for identifying new chemical entities

The number of applications of VS is rapidly escalating in recent times. Several successful cases have been reported which resorted to the synthesis of lead compounds by VS methods for various targets. In this article, we have tried to provide an overview of these studies in **Table 2**.

6. Databases for VS

One needs to remember that the database library must fit the purpose of the experiment before its selection for screening. **Table 3** summarizes a representative list of public and commercial chemical databases that are commonly screened in real practices.

7. Expert opinion

VS approaches have been dynamically adopted by pharmaceutical companies with intent to obtain as many potential compounds as possible hoping for the greater chance of finding hits from chemical libraries. Many successful examples have been demonstrated in recent years in the field of computer-aided VS for lead identification, using receptor-based or ligand-based approaches. However, the initial excitement has collapsed considering the less than desired outcomes from these combinatorial screening methods. There appears to be no universal method to carry out these studies as each biological target system is unique in nature. What has led to these successes is an in-depth understanding of the target under investigation and fine-tuning of the VS scheme to achieve the desired result. In most cases, this was done by applying all the available information to generate and validate the models. It wisely includes the course of database screening with known actives and then defining the workflow so as to identify the known ligands as high ranking hits. Some of the key challenges in VS are the appropriate treatment of ionization, tautomerization of ligand and protein residues, target/ligand flexibility, choice of force fields, solvation effects, dielectric constants, exploration of multiple binding modes, consideration of water molecules in proteins and, most importantly, the approximations in the scoring functions that lead to false-positives and miss true-hits.

Though one cannot ignore the inherent limitations of VS, still it is one of the best options now available to explore a large chemical space in terms of cost-effectiveness and commitment in time and material. It allows access to a large number of possible ligands to explore and most importantly many of them are easily available for purchase and subsequent test. The number of therapeutic targets that have been fully characterized by crystallography is currently limited but this situation is set to significantly change in the immediate future as structural genomics and proteomics initiatives begin to yield fruitful results. With the development of new docking methodologies, ligand-based screening techniques and machine learning tools, VS techniques are capable of predicting better hit rates. VS methodologies will play front runner role in drug design in a near future either as a complementary approach to HTS or as standalone approach. In the opinion of the present authors, the technologies are there and they just need to be employed in right way and in right direction to identify novel new chemical entities with the scientific utilization of the VS techniques. In consequence, the newer, safer and effective drug innovation will be a matter of time and dedicated research continues keeping our hope for blockbuster drugs alive.

Declaration of interest

S Kar thanks the Department of Science and Technology (DST), Government of India for awarding a Research fellowship under the INSPIRE scheme. Council of Scientific and Industrial Research (CSIR), New Delhi and Department of Biotechnology (DBT), Government of India, New Delhi are also thanked for awarding major research projects to KR.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Shoichet BK. Virtual screening of chemical libraries. Nature 2004;432:862-5
2. Jahn A, Hinselmann G, Fechner N, et al. Optimal assignment methods for ligand-based virtual screening. J Cheminform 2009;1:14

• **Important in terms of ligand-based VS.**

3. Villoutreix BO, Renault N, Lagorce D, et al. Free Resources to Assist Structure-Based Virtual Ligand Screening Experiments. Curr Protein Pept Sci 2007;8:381-411

• **Important in terms of structure-based VS.**

4. Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discov 2002;1:882-94

•• **Important in terms of VS methods.**

5. Reddy AS, Pati SP, Kumar PP, et al. Virtual Screening in Drug Discovery-A Computational Perspective. Curr Protein Pept Sci 2007;8:329-51
6. Cheng T, Li Q, Zhou Z, et al. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. AAPS J 2012;14:133-41
7. Muegge I, Oloff S. Advances in virtual screening. Drug Discov Today 2006;3:405-11
8. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl 2002;41:2644-76
9. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual Screening. Drug Discov Today 2011;16:372-6
10. Sun H. Pharmacophore-based virtual screening. Curr Med Chem 2008;15:1018-24
11. Schneider G, Böhm H. Virtual screening and fast automated docking methods: combinatorial chemistry. Drug Discov Today 2002;7:64-70
12. Waszkowycz B, Perkins T, Sykes R, et al. Large-scale virtual screening for discovering leads in the postgenomic Era. IBM Syst J 2001;40:360-76
13. Reddy ChS, Vijayasarathy K, Srinivas E, et al. Homology modeling of membrane proteins: a critical assessment. Comput Biol Chem 2006;30:120-6
14. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 2003;9:47-57
15. Scior T, Bender A, Tresadern G, et al. Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 2012;52:867-81

•• **Important in terms of pitfalls in VS methods.**

Affiliation

Supratik Kar & Kunal Roy[†]
[†]Author for correspondence
Jadavpur University,
Drug Theoretics and Cheminformatics Laboratory,
Department of Pharmaceutical Technology,
Kolkata 700032, India
Tel: +91 98315 94140;
Fax: +91 33 2837 1078;
E-mail: kunalroy_in@yahoo.com