



ELSEVIER

# Integrating virtual screening in lead discovery

Tudor I Oprea<sup>a,\*</sup> and Hans Matter<sup>b</sup>

Target- and ligand-based virtual screening have emerged as resource-saving techniques that have been successfully applied to identify novel chemotypes in biologically active molecules. Eight confirmed virtual screening hits have recently been described and are discussed in this review, with focus on the workflow. These are then evaluated in the light of pharmacokinetics prediction (e.g. Caco-2 permeability, cytochrome P450 inhibition and hERG binding). We anticipate problems for five of these hits (e.g. cardiac toxicity), which warrant further experiments. Future challenges include dynamic tautomer/protonation treatment for both ligands and targets and improved pre- and post- virtual screening filters.

## Addresses

<sup>a</sup>Division of Biocomputing, University of New Mexico School of Medicine, MSC 08 4560, 1 University of New Mexico, Albuquerque, New Mexico 87131-0001, USA

\*correspondence: toprea@salud.unm.edu

<sup>b</sup>Aventis Pharma Deutschland GmbH, DI&A Chemistry, Building G 878, D-65926, Frankfurt am Main, Germany  
e-mail: hans.matter@aventis.com

**Current Opinion in Chemical Biology** 2004, **8**:349–358

This review comes from a themed issue on  
Next-generation therapeutics  
Edited by Tudor Oprea and John Tallarico

Available online 2nd July 2004

1367-5931/\$ – see front matter  
© 2004 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.cbpa.2004.06.008

## Abbreviations

<b>ADMET</b>	absorption, distribution, metabolism, excretion and toxicity
<b>BBB</b>	blood–brain barrier
<b>CATS</b>	chemically advanced template search
<b>CK2</b>	casein kinase II
<b>ER<math>\alpha</math></b>	estrogen receptor $\alpha$
<b>hERG</b>	human ether-a-go-go related gene
<b>LBVS</b>	ligand-based virtual screening
<b>PDB</b>	Protein Data Bank
<b>P-gp</b>	P-glycoprotein
<b>PPB</b>	plasma protein binding
<b>QSAR</b>	quantitative structure–activity relationship
<b>TBVS</b>	target-based virtual screening
<b>TGT</b>	tRNA-guanine transglycosylase
<b>TR</b>	thyroid hormone receptor
<b>U-II</b>	urotensin II
<b>VCAM-1</b>	vascular cell adhesion molecule-1
<b>VDss</b>	volume of distribution at steady state
<b>VS</b>	virtual screening

## Introduction

Economic pressure to deliver ‘best-in-class’ drugs on the market, which partly explains the ‘innovation deficit’ [1]

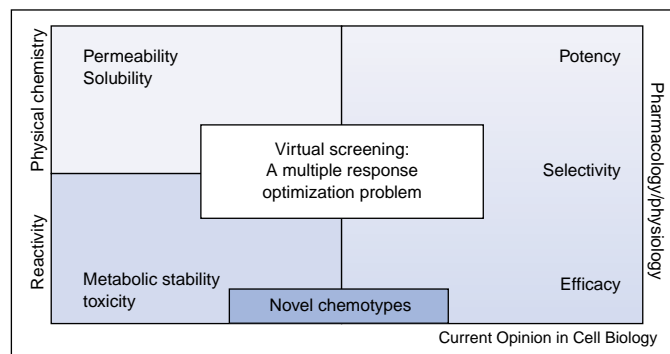
of the pharmaceutical industry, forces drug discovery scientists to develop computational alternatives that lead to the identification of isosteric molecular scaffolds that move beyond the sampling of local chemical spaces. Computational chemistry has partially fulfilled its promise [2], but the reality of drug–receptor interactions, at the molecular level, continues to be too complex to provide a failsafe *in silico* technology for drug discovery [3]: entropy and the dielectric constant are just two examples of subjects continuously debated among experts. The challenges of *in silico* drug discovery include the evaluation of multiple binding modes, of accessible conformational states for both ligand and receptor, of affinity and selectivity versus efficacy, plasma protein binding, metabolic stability (site of reactivity and turn-over), absorption, distribution and excretion, as well as *in vivo* versus *in vitro* properties of model compounds, while seeking a favorable intellectual property position (see also Figure 1). These difficulties are hardly surmountable by any single computer-aided molecular design software to date.

The key decision step of selecting higher quality compounds has shifted from the drug candidate to the lead optimization and even lead identification stages [4], and we are witnessing an increased integration of *in silico* technologies that sift through enormous numbers of virtual chemicals based on ‘soft modeling’ quantitative structure–activity relationship (QSAR) techniques [5] that estimate properties deemed important for orally available drugs (Figure 1). The ‘fail early’ strategy (i.e. the effort to terminate projects as early as possible, before significant investments are made in a given project) places a significant responsibility, and thus an increased demand in prediction quality for all virtual screening (VS) technologies. Recently established [6,7], VS is regarded as a complement to bioactivity screening [8]. For cases where structural information about the target is available from either fact or inference, we can consider target-based virtual screening (TBVS); all other cases can be defined as ligand-based virtual screening (LBVS). VS technologies appeared in 1997 and became mainstream after 2000 (Figure 2). A literature search shows 211 CAPLUS entries in 2002 and 2003, of which 34 appear to indicate some ‘success’. This paper briefly outlines the TBVS and LBVS methodologies, highlighting eight of the successful VS applications, with focus on the workflow and pharmacokinetics, or ADMET (absorption, distribution, metabolism, excretion and toxicity) evaluation. We then discuss future challenges.

## Target-based virtual screening

Given a 3D structure of the target, TBVS relies on docking and scoring to provide potential candidates for further analysis [9]. AutoDock [10], DOCK [11], FlexX

Figure 1



Virtual screening, seen as an integrative approach to address a multiple response surface optimization problem. Properties related to physical chemistry, to chemical reactivity, to pharmacology and physiology require optimization for novel chemotypes, should the outcome result in a launched drug. The increased difficulty of finding such optima is suggested by darker backgrounds.

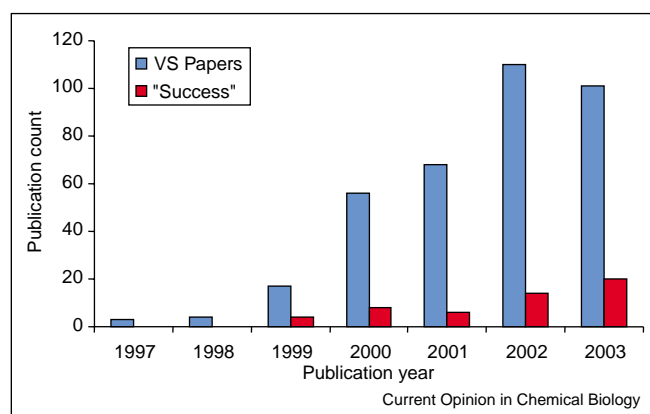
[12], FRED [13], GOLD [14] and ICM [15] are some of the most used docking programs in the field. Most of these programs can dock libraries of single structures or even multiconformer libraries, and are thus suitable for high-throughput searches. Regardless of the choice of the docking software, one is limited in TBVS by the ability to score the ligands in an appropriate manner. There are four major categories of scoring functions. First, knowledge-based methods [16–18] use Boltzmann-weighted potentials of mean force derived from statistical analyses of ligand–receptor inter-atomic contacts, based on available complexes in the Protein Data Bank (PDB) [19]. SMOG [20], Muegge's [21], and Drug-Score [22] are implementations of this approach. Second, 'master equation' approaches [23] estimate the energetic contributions of various interaction types in a semiquantitative manner [24–26]. Third, regression-based [27,28]

methods take advantage of the available biological activity for training sets of ligand–receptor complexes extracted from the PDB [29–32]. Finally, Poisson–Boltzmann equation solvers that address electrostatics incorporating solvent effects [33]. Consensus scoring [34] can be used when a particular scheme fails. We recently showed that pharmacokinetic awareness [3] can be integrated into a regression-based scoring scheme [35], using VolSurf [36].

### Ligand-based virtual screening

Given a bioactive (rigid) conformer derived from structural methods (X-ray and NMR), or from molecular modeling, LBVS can rank novel ligands by 3D similarity searching or by pharmacophore pattern matching. Both methods rely on appropriate software to query large databases of virtual or existing chemicals.

Figure 2



Histogram analysis of VS publications indexed by the Chemical Abstracts Service between 1997 and 2003. Database used, CAPLUS; search date, March 19, 2004; search limited to 2003; keywords, 'virtual screening' – 359 entries (blue bars), 'success' in this subset returned 52 entries (red bars). SciFinder Scholar 2002, © American Chemical Society.

### Pharmacophore identification and matching

In the pharmacophore concept [37], all ligands, regardless of chemotype, present similar steric and electrostatic features that are recognized at the target binding site and are responsible for the biological activity. Built on Marshall's active analog approach [38], pharmacophore perception methods — discussed in a book edited by Güner [39] — apply several criteria for validation and searching. Known actives include at least one rigid representative, which becomes the template structure, and pharmacophore features must apply to at least one (class of) conformer(s) from each active ligand, while negative features (e.g. restricted steric volumes) should apply mostly to inactive molecules. Some methods, for example ALMOND [40], correlate pharmacophoric patterns with biologic activity using statistical methods.

### 3D-shape and similarity searching

In this paradigm, molecules with similar features have similar biologic activity [41]. Tanimoto's 'distance-between-patterns' [42] (symmetric) and Tversky's asymmetric 'contrast model' [43] are the basis for similarity measure [44–47]. Chemical similarity relates to the framework of a descriptor system and to that of an object or class of objects [48], and depends on the choice of molecular descriptors [49], the choice of the weighting scheme(s) and on the similarity coefficient itself. In 3D-similarity LBVS, the database queries start compare molecules (steric, electrostatic similarity) with the 3D information derived from known actives (derived from fact or inference). ROCS (rapid overlay of chemical structures), a shape-based superposition and database search tool from OpenEye (<http://www.eyesopen.com>), uses a Gaussian representation of the molecular volume [50] and has

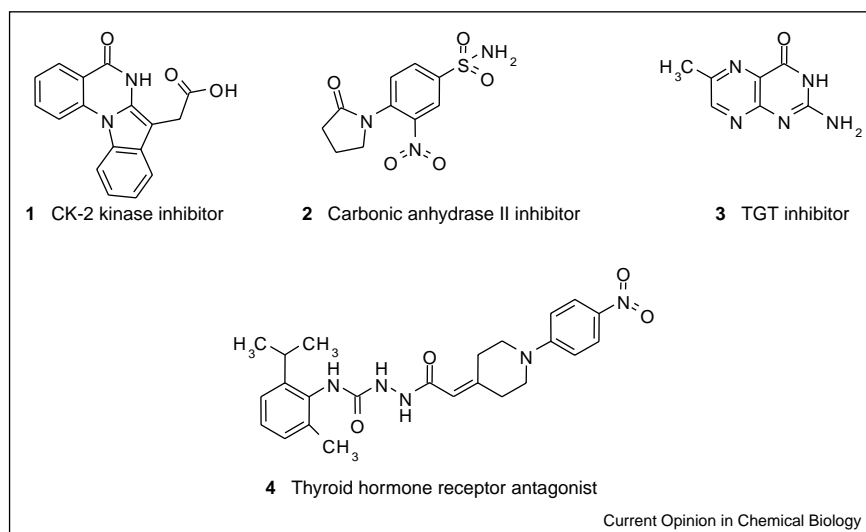
several options (e.g. chemical fragment weighted force-fields) that merge elements of 2D-similarity into LBVS.

### Successful applications of target-based virtual screening

Novel, potent and selective CK2 (casein kinase II) inhibitors [51•] were obtained by screening a subset (ca. 400 000 molecules) of the Novartis database using DOCK [11]. The ATP binding site of a human CK2 $\alpha$  was inferred via homology modeling (X-ray structure of *Zea mays* CK2 $\alpha$ , PDB entry 1DAW). Filters related to molecular weight, to the number of rotatable bonds, and to undesired substructures were applied before TBVS. Protonation states were adjusted using a rule-based method. The compounds with best DOCK scores were further refined using known pharmacophore information (e.g. hydrogen bonds to the kinase hinge region). All acceptable compounds were re-evaluated using SCORE [52]. The best compounds were finally inspected before submitting a total of 12 compounds for biologic screening. The best rigid inhibitor has an IC<sub>50</sub> of 80 nM (Compound 1, Figure 3). Novel BCR-ABL tyrosine kinase inhibitors were identified with a related TBVS workflow [53], using DOCK [11] and a database of 200 000 commercially available compounds. The top 1000 candidates, as ranked by the DOCK energy score, were analyzed before selecting 15 structurally diverse compounds. Eight of these show IC<sub>50</sub> values of 10–200  $\mu$ M.

Human carbonic anhydrase II inhibitors were identified [54•] using a set of hierarchical filters and the FlexX docking engine [12], applied to a database of ~100 000 compounds. The TBVS workflow started with a pre-selection based on privileged functional groups from an

Figure 3



Chemical structures for successful TBVS runs, categorized by target; see text for details.

analysis of available ligand and protein structural data, followed by pharmacophore searching using a pharmacophore derived from a protein binding site. The similarity of candidates with known ligands was used to re-rank the VS hitlist from this step. The top 100 compounds after this step were subject to flexible docking using FlexX. After affinity prediction analysis, 13 compounds were tested. Three molecules exhibited subnanomolar range affinities (e.g. **2**), as well as novel chemotypes. Furthermore, the binding modes inferred from docking were experimentally validated by crystallizing two of these inhibitors. A similar workflow, but starting with an initial database of 800 000 molecules, was applied [55] to identify lead structures for tRNA-guanine transglycosylase (TGT) based on an X-ray structure analysis of a previous micromolar inhibitor. Given its unexpected binding mode, three differing target-based pharmacophore hypotheses were generated and employed in TBVS. The most acceptable compounds were docked into one of the two alternating TGT binding site conformers using FlexX. Of the final set of nine molecules with micromolar and submicromolar activities, one commercially available structure ( $IC_{50} = 250$  nM) is shown (**3**).

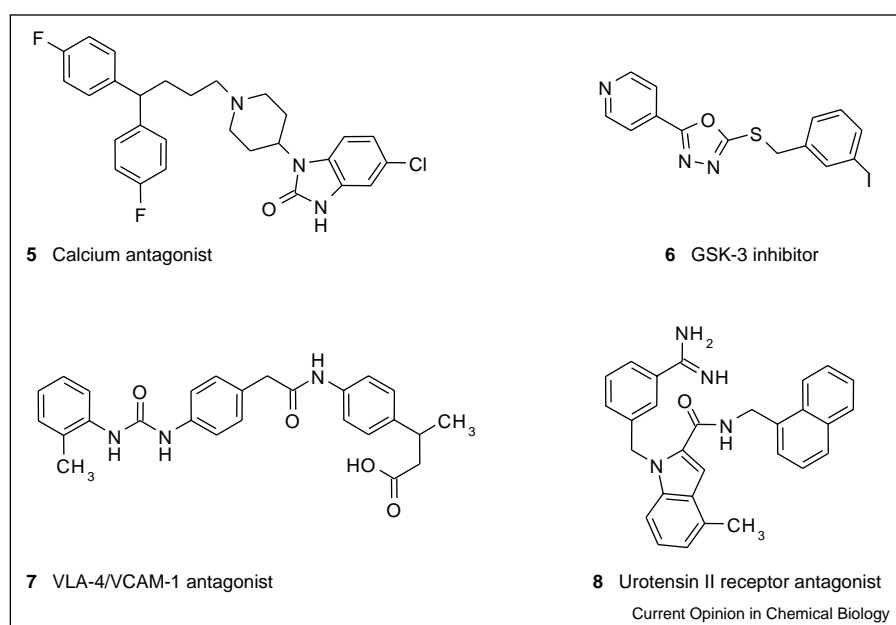
When screening nuclear hormone receptor targets for antagonists, one typically starts by inferring the antagonist-binding conformation of the target from one of the antagonist-bound estrogen receptor  $\alpha$  (ER $\alpha$ ) structures. ICM [15] was used in a TBVS study [56] to identify thyroid hormone receptor (TR) antagonists, starting from the antagonist-bound conformation of TR, built from the raloxifene-bound ER $\alpha$  (PDB code 1ERR). The TBVS

workflow started with the Lipinski's 'rule of five' [57] filtering on a database of ca. 250 000 structures, followed by rigid ICM docking. Because known ER $\alpha$  antagonists exhibit steric bulk in a certain receptor area (helix H12), only the top 1000 ligands that had exhibited this type of steric bulk were examined further. After flexible receptor/ligand ICM docking, the top 300 hits were manually inspected, and a set of 100 candidates was subjectively selected for biologic screening. Fourteen of 75 tested molecules were found to antagonize this receptor with  $IC_{50}$  values ranging from 1.5 to 30  $\mu$ M. One of the best ligands from this run was used to design a follow-up virtual library. After synthesis, some of these second generation antagonists showed submicromolar activities (e.g. **4**). Additional TBVS successes are discussed in this issue [58].

### Successful applications of ligand-based virtual screening

CATS (chemically advanced template search) [59], a topologically-based pharmacophore screening method, has the ability to perform 'scaffold-hopping' (i.e. to identify isofunctional structures that differ in chemotype). When using mibefradil, a known T-type calcium channel blocker ( $IC_{50} = 1.7$   $\mu$ M) as a query, CATS identified one significant hit (clopimozid; **5**, Figure 4) with an  $IC_{50} < 1$   $\mu$ M, among the top 12 ranked molecules. The same technique, CATS [59], was applied to identify structurally novel glycogen synthase kinase-3 inhibitors [60], first by identifying the oxadiazol-pyridyl moiety, a new chemotype, then by synthesizing additional analogs. Compounds with inhibitory activity below 1  $\mu$ M were identified (e.g. **6**).

Figure 4



Chemical structures for successful LBVS runs, categorized by target; see text for details.

Pharmacophore-searching techniques have also resulted in several successful LBVS applications. Potent and novel integrin  $\alpha 4\beta 1$  antagonists were identified [61<sup>\*</sup>] using Catalyst [62], employing a pharmacophore query obtained from a tetrapeptide inhibitor. The fibronectin-derived semi-synthetic peptide, a strong binder to  $\alpha 4\beta 1$  ( $IC_{50} = 0.6$  nM), was modeled to match the X-ray structure of the related integrin-binding region of vascular cell adhesion molecule-1 (VCAM-1, PDB entry code 1VCA). The computational screen identified 12 matches from a virtual library of 8624 molecules that satisfied the model and additional synthetic filters. All synthesized molecules were found to inhibit the interaction between  $\alpha 4\beta 1$  and VCAM-1, with the most potent candidate showing an  $IC_{50}$  value of 1.3 nM, comparable to the starting compound (7).

A similar LBVS workflow was followed to identify non-peptidic urotensin II (U-II) receptor antagonists, using a G-protein coupled receptor-biased subset of the Aventis database [63<sup>\*</sup>]. Based on an alanine-scan of U-II that resulted in 25 peptide analogs, combined with NMR studies of U-II and of an active peptide, a pharmacophore model was derived from the NMR-derived solution structure of U-II, using three pharmacophore elements: Trp-7 and Tyr-9 (hydrophobes), and Lys-8 (positive charge). In addition to these three features, the shape of the message sequence was included in the pharmacophore. Subsequent VS returned six different scaffold classes, with the most active U-II receptor antagonist showing an  $IC_{50}$  value of 400 nM (8).

### Virtual screening hits: a critical pharmacokinetics evaluation

The eight papers reviewed above highlight issues related to higher binding affinity and chemotype novelty, when presenting their VS workflow and success. Although some of these authors have industrial affiliation or connections, there is no discussion of measured or inferred pharmacokinetic properties of these compounds. Because this and another recent review on computer-aided drug discovery [64] advocate the *in silico* evaluation of VS hits for pharmacokinetic properties, we discuss the above eight struc-

tures (designated by their numbers in Table 1) in the light of *in silico* evaluations performed with software available from Molecular Discovery (<http://www.moldiscovery.com>). We are aware that there are different software tools for ADMET prediction, but it is beyond the scope of this review to discuss this area (see this issue for a review on this topic [65]). ADMET prediction for these VS hits is based only on models derived using three programs (MetaSite, VolSurf and Almond – see below), but a consensus approach might increase the reliability of predictions. Software and models for *in silico* profiling are usually chosen based on internal experience (e.g. prediction reliability for in-house data), consistency of the approach and its integration into current drug discovery workflows. The early *in silico* profiling for liabilities is intended to increase awareness (i.e. to provide a ‘computational alert’ [57]) regarding issues that might require attention in later (e.g. lead optimization) stages. They require experimental profiling and appropriate adjustment of models to represent the chemotype of interest, and are intended to prioritize experimental procedures that would rule out (or confirm) these problems — to avoid late-phase failures. Therefore, all computational results summarized in Table 1 should be regarded as *potential*, not actual problems for these compounds. Each property evaluated *in silico* for the eight VS hits and summarized in Table 1 is briefly outlined below:

### Cytochrome P450 (CYP) inhibition and substrate prediction

Binding for the two most important isozymes in drug metabolism (CYP2D6 and CYP3A4) was evaluated with Metasite 1.8, using over 100 known substrates and inhibitors for each isozyme, as well as 3D models built from crystallographic data [66] using homology modeling. A similar procedure was described for CYP2C9 [67,68<sup>\*</sup>]. Substrate hydrogens are ranked in terms of site of metabolism probability [69] using heme Fe proximity, while a docking-based procedure evaluates the ability of potential inhibitors to bind anywhere in the CYP binding site (G Cruciani, personal communication). In Table 1, results can be interpreted as follows: ‘potent’ implies that the

Table 1

Pharmacokinetic property predictions for the molecules depicted in Figures 3 and 4; see text for details.

ID	Target	$IC_{50}$ (nM)	CYP2D6 inhibitor	CYP3A4 inhibitor	P-gp substr.	BBB perm.	Caco-2 perm.	Water sol.	DMSO sol.	PPB (%)	VDss (l/kg)	hERG binding
1	CK-2	80	Weak	Potent	No	n.p.	n.p.	Yes	n.p.	60.91	0.9	Potent
2	CA-II	0.6	Weak	Potent	No	No	n.p.	n.p.	No	37.7	0.7	Potent
3	TGT	2700	Weak	Weak	No	No	n.p.	Less	No	38.98	0.3	Potent
4	TR	750	Weak	Weak	n.p.	No	Yes	Yes	Yes	92.85	2.9	Potent
5	T-channel	<1000	Potent	n.p.	Yes	n.p.	Yes	Yes	Yes	99	6.6	V. weak
6	GSK-3	390	Weak	Potent	Yes	n.p.	Yes	Yes	Yes	86.78	1.7	Weak
7	VLA-4	1	Weak	Weak	n.p.	No	n.p.	Yes	Less	83.7	3	Potent
8	U-II rec.	400	Weak	Weak	n.p.	Yes	Yes	Yes	Yes	99	8.9	V. weak

DMSO, dimethyl sulfoxide; n.p., not predicted.

compound is predicted to have a binding affinity to that isozyme of less than 1  $\mu\text{M}$ , i.e. it is expected to be a relatively potent inhibitor, while ‘weak’ implies inhibition occurs in the micromolar range. For this, and all other predictions in Table 1, ‘n.p.’ (not predicted) refers to a high degree of ambiguity regarding the prediction – thus, no prediction is offered. Four of the 8 VS hits might be potent inhibitors: (5). for CYP2D6, and 1, 2 and 6 for CYP3A4. None of the compounds were predicted to be substrates.

#### P-glycoprotein substrate predictions

These were performed using Almond 3.2, pharmacophore pattern analysis software that combines molecular interaction field distances with energies [40], using a model initially derived from P-glycoprotein (P-gp) ATPase activity [70]. This model contains 100 drugs (60 known substrates, and 40 non-binders), each evaluated over 100 diverse conformations (G. Cruciani, personal communication). Two of the VS hits (5 and 6) are estimated to be P-gp substrates, three molecules are ambiguous, and another three are estimated not to be substrates.

#### Blood–brain barrier permeability

This and all the remaining predictions in Table 1 were performed using VolSurf 4.0 [36], software that condenses three-dimensional molecular interaction field information into two-dimensional descriptors. VolSurf applications to drug discovery were recently reviewed [71<sup>\*</sup>]. Blood–brain barrier (BBB) permeability is predicted from a discriminant analysis model [72]. Four of the VS hits appear not to permeate the brain (2, 4 and 7); (8) is predicted to penetrate, while the prediction for three molecules is uncertain.

#### Caco-2 permeability prediction

This was based on an experimental model (Caco-2 cells monolayer) that evaluates the intestinal absorption of drugs [73], and was derived from known literature datasets (see [74] for a comprehensive review). We estimate that four out of eight VS hits will cross the intestinal barrier, while the prediction for four molecules (1, 3 and 7) is ambiguous. Molecule 8 contains a highly basic, permanently charged benzamidinium group, but VolSurf handles only neutral molecules. Thus, VolSurf predictions for Caco-2 and BBB permeability (as well as solubility and other properties) for charged species should be treated with care.

#### Water solubility prediction

This is based on various literature datasets [75,76], and is comparable to other models [71<sup>\*</sup>]. Except for compounds 2 (ambiguous) and 3 (less soluble), all VS hits are predicted to be suitable for development.

#### DMSO solubility prediction

This is based on experimental determinations from the University of Perugia [77]. Four out of the eight VS hits appear to have suitable DMSO solubility, while com-

pounds 2 and 3 are predicted to be insoluble. Compound 1 is ambiguous, while 7 appears to be less soluble.

#### Plasma protein binding prediction

This is based on collected plasma protein binding (PPB) percentage values for therapeutic drugs from the literature [78<sup>\*</sup>]. In this VolSurf model, values predicted to be 95% or higher are equivalent (due to model inaccuracy for high PPB values), whereas lower values (a desired feature in drugs) are more accurate [71<sup>\*</sup>]. Only two out of eight VS hits (5 and 8) are expected to show high percentage PPB, whereas the other six appear to be suitable for further development.

#### Volume of distribution

This is a drug-disposition parameter relating the amount of drug in the body to the concentration of the drug in the blood (or tissue), which tries to address the ‘how often’ question in the therapeutic dose regimen. The  $VD_{ss}$  (volume of distribution at steady state, the most commonly accepted abbreviation for volume distribution in pharmacokinetics literature) prediction model is similar to that from Pfizer [79,80]. Three compounds (1, 3) have good  $VD_{ss}$  values, three have marginally good values (4, 6 and 7), while two might pose problems later on (5 and 8).

#### Human ether-a-go-go related gene

Human ether-a-go-go related gene (hERG) encodes a  $K^+$  channel that is implicated in the fatal arrhythmia known as *torsade de pointes*. It appears to be the molecular target responsible for the cardiac toxicity of a wide range of therapeutic drugs [81]. The hERG binding VolSurf model is based on over 200 drugs collected from literature [82<sup>\*</sup>,83,84], and is similar to work performed at Roche [85<sup>\*\*</sup>]. On the basis of this model, it appears that only two molecules, 5 (a  $Ca^{2+}$ -channel blocker) and 8, are very weak hERG binders; one is a weak binder 6; while five compounds (1, 4 and 7) are flagged for possible cardiac toxicity.

The above property forecasts are given as an example of the possible applications of multiple response optimization. hERG binding is the most critical of these properties because potent binding to this  $K^+$  channel may be lethal [81]. The next most important property is CYP2D6 inhibition, because this isozyme does not appear to be inducible and yet is responsible for metabolizing one-fifth of known drugs. Only three VS hits (6, 7 and perhaps 5) appear to satisfy these critical properties, but P-gp efflux may have to be evaluated for two of these compounds (5 and 6).

The quality of ADMET predictions relates to our ability to estimate highly charged molecules (e.g. *in vitro* and *in vivo* studies show that permanently charged molecules have a very low tendency to permeate across intestinal and other biological barriers via passive diffusion). This

behaviour is not properly estimated by *in silico* models. Similar observations have been made using current software to predict solubility for charged molecules. We expect advances in this area, as more tailored ADMET prediction software that gives proper treatment to charged functional groups will yield improved predictivity. Consensus predictions, using different software tools, might highlight such discrepancies. Since the success rate for predictions is not 100%, we do not exclude any of the VS hits from further considerations after *in silico* profiling. Instead, we propose to experimentally test for potential liabilities early in the decision tree and use this knowledge to potentially build a more specific model during lead optimization.

### Conclusions and future challenges

There is no general solution for TBVS or LBVS. Every prospective application requires understanding and tuning of key parameters. The user should apply all available information to generate and validate the models (e.g. binding site in TBVS, bioactive conformation in LBVS), then define the workflow accordingly. Filters of increasing complexity (and reduced computational speed) should be applied as the VS proceeds [86]. Filters can significantly reduce the number of candidates for screening by eliminating structures with undesired chemotypes, by applying limits to certain physico-chemical properties (e.g. rule-of-five [57]), or criteria derived by observing the property distribution in drug-related databases [87].

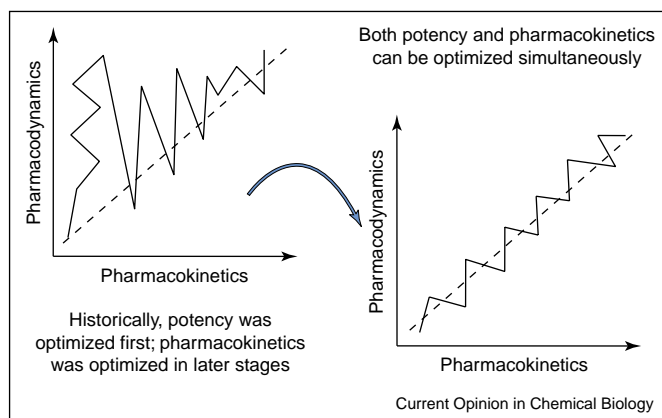
TBVS campaigns require database filtering and target preparation, pre-docking filters, docking, post-docking analysis and prioritization of compounds for testing [88]. LBVS campaigns require careful validation of the pharmacophore/3D-shape pattern, seamless link to 3D-database searching, rapid feature detection and the inclusion of sterically forbidden regions, access to multiple

conformers per molecule and multiple models (patterns) evaluation [89\*].

A key challenge in VS is the appropriate treatment of ionization and tautomerization states in the input data. There is currently no software that performs the required corrections in an automatic fashion, although OpenEye has announced plans for incorporating tools for static protein  $pK_a$  and dynamic protein  $pK_a$  prediction, as well as for ligand tautomer enumeration [90]. Docking the correct ligand tautomer requires dynamic protein  $pK_a$  prediction, since tautomers are influenced by environment. This leaves open the issue of tautomerism during LBVS. Because the ligands are evaluated in 'gas phase' during pharmacophore or 3D-shape analyses, the user may have to enumerate tautomers and ions for each of the screening candidates. Other challenges in the field relate to ligand/target flexibility, the choice of force-fields, partial charges, solvent, dielectric constant and to the exploration of multiple binding modes.

The greatest challenge remains the simultaneous optimization of both binding affinity and pharmacokinetic properties [3] (see Figures 1,5). The difficulty consists in addressing the appropriate molecular determinants that define the desired compound characteristics, in a consistent manner. For example, both hydrophobicity and hydrogen bonds contribute to binding affinity and to passive permeability. The question remains, in what proportion are they contributing to each property? One solution is the development of an integrated software framework that monitors ligand (or library) alterations in the context of 'fitness landscape' (via docking and scoring, or perhaps via 3D-shape/pharmacophore matching). We have illustrated the incorporation of VolSurf — a pharmacokinetics prediction tool *par excellence*, into a TBVS environment [35\*] that is amenable to high-throughput pharmacokinetics prediction [91]. It is conceivable that,

Figure 5



The paradigm shift in lead discovery oriented VS: Reducing the number of evaluation steps while finding an optimum in multiple property spaces. Modified from [3] with permission.

in the next decade, such integrated methods will become mainstream.

## Acknowledgements

We thank Gabriele Cruciani (University of Perugia, Italy) and Ismael Zamora (Lead Molecular Design, Barcelona, Spain) for discussions regarding pharmacokinetic property predictions. This work was supported in part by New Mexico Tobacco Settlement funds (TIO).

## References and recommended reading

Papers of particular interest, republished within the annual period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Drews J: **Innovation deficit revisited: reflections on the productivity of pharmaceutical R&D.** *Drug Discov Today* 1998, **3**:491-494.
  2. Horrobin DF: **Innovation in the pharmaceutical industry.** *J R Soc Med* 2000, **93**:341-345.
  3. Oprea TI: **Virtual screening in lead discovery: A viewpoint.** *Mol* 2002, **7**:51-62.
  4. Oprea TI: **Lead structure searching: Are we looking at the appropriate property?** *J Comput Aided Mol Des* 2002, **16**:325-334.
  5. Stone M, Jonathan P: **Statistical thinking and technique for QSAR and related studies. Part II: Specific methods.** *J Chemomet* 1994, **8**:1-20.
  6. Horvath D: **A virtual screening approach applied to the search for trypanothione reductase inhibitors.** *J Med Chem* 1997, **40**:2412-2423.
  7. Walters WP, Stahl MT, Murcko MA: **Virtual screening - an overview.** *Drug Discov Today* 1998, **3**:160-178.
  8. Mestres J: **Virtual screening: a real screening complement to high-throughput screening.** *Biochem Soc Trans* 2002, **30**:797-799.
  9. Muegge I, Enyedy I: **Docking and Scoring.** In *Computational Medicinal Chemistry and Drug Discovery*. Edited by Tollenaere J, De Winter H, Langenaeker W, Bultinck P. New York: Marcel Dekker; 2004:405-436.
  - Comprehensive review of docking and scoring methods in drug discovery.
  10. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS: **Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in Autodock.** *Proteins Struct Funct Genet* 2001, **46**:34-40.
  11. Ewing TJA, Makino S, Skillman AG, Kuntz ID: **DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases.** *J Comput Aided Mol Des* 2001, **15**:411-428.
  12. Kramer B, Metz G, Rarey M, Lengauer T: **Ligand docking and screening with FlexX.** *Med Chem Res* 1999, **9**:463-478.
  13. McGann M, Almond H, Nicholls A, Grant JA, Brown F: **Gaussian docking functions.** *Biopolymers* 2003, **68**:76-90.
  14. Jones G, Willett P, Glen RC, Leach AR, Taylor R: **Development and validation of a genetic algorithm for flexible docking.** *J Mol Biol* 1997, **267**:727-748.
  15. Totrov M, Abagyan R: **Flexible protein-ligand docking by global energy optimization in internal coordinates.** *Proteins Struct Funct Genet* 1998, (Suppl. 1):215-220.
  16. Sippl MJ: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *J Comput Aided Mol Des* 1993, **7**:473-501.
  17. Sippl MJ: **Knowledge-based potentials for proteins.** *Curr Opin Struct Biol* 1995, **5**:229-235.
  18. Domingues FS, Koppensteiner WA, Jaritz M, Prlic A, Weichenberger C, Wiederstein M, Floeckner H, Lackner P, Sippl MJ: **Sustained performance of knowledge-based potentials in fold recognition.** *Proteins Struct Funct Genet* 1999, (Suppl 3):112-120.
  19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
  20. Ishchenko AV, Shakhnovich EI: **Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions.** *J Med Chem* 2002, **45**:2770-2780.
  21. Muegge I, Martin YC: **A general and fast scoring function for protein-ligand interactions: A simplified potential approach.** *J Med Chem* 1999, **42**:791-804.
  22. Sotriffer CA, Gohlke H, Klebe G: **Docking into knowledge-based potential fields: a comparative evaluation of drugscore.** *J Med Chem* 2002, **45**:1967-1970.
  23. Ajay, Murcko M: **Computational methods to predict binding free energy in ligand-receptor complexes.** *J Med Chem* 1995, **38**:4953-4967.
  24. Williams DH, Cox JPL, Doig AJ, Gardner M, Gerhard U, Kaye PT, Lai AR, Nicholls IA, Salter CJ, Mitchell RC: **Toward the semiquantitative estimation of binding constants. Guides for peptide-peptide binding in aqueous solution.** *J Am Chem Soc* 1991, **113**:7020-7030.
  25. Williams DH, Bardsley B: **Estimating binding constants - The hydrophobic effect and cooperativity.** *Perspect Drug Discov Des* 1999, **17**:43-59.
  26. Marrone TJ, Luty BA, Rose PW: **Discovering high-affinity ligands from the computationally predicted structures and affinities of small molecules bound to a target: A virtual screening approach.** *Perspect Drug Discov Des* 2000, **20**:209-230.
  27. Böhm HJ: **The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure.** *J Comput Aided Mol Des* 1994, **8**:243-256.
  28. Verkhivker G, Appelt K, Freer ST, Villafranca JE: **Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of HIV-1 protease binding affinity.** *Protein Eng* 1995, **8**:677-691.
  29. Böhm HJ: **Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs.** *J Comput Aided Mol Des* 1998, **12**:309-323.
  30. Head RD, Smythe ML, Oprea TI, Waller CL, Greene SM, Marshall GR: **VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands.** *J Am Chem Soc* 1996, **118**:3959-3969.
  31. Marshall GR, Head RD, Ragno R: **Affinity prediction: The sine qua non.** In *Thermodynamics in Biology*. Edited by Di Cera E. New York: Oxford University Press; 2001:87-111.
  32. Wang R, Lai L, Wang S: **Further development and validation of empirical scoring functions for structure-based binding affinity prediction.** *J Comput Aided Mol Des* 2002, **16**:11-26.
  33. Grant JA, Pickup BT, Nicholls A: **A smooth permittivity function for Poisson-Boltzmann solvation methods.** *J Comput Chem* 2001, **22**:608-640.
  34. Wang R, Wang S: **How does consensus scoring work for virtual library screening? An idealized computer experiment.** *J Chem Inf Comput Sci* 2001, **41**:1422-1426.
  35. Zamora I, Oprea TI, Cruciani G, Pastor M, Ungell AL: **Surface descriptors for protein-ligand affinity prediction.** *J Med Chem* 2003, **46**:25-33.
  - Promising integration of affinity and ADMET prediction at the level of the employed descriptors as important prerequisite for subsequent lead optimization.
  36. Cruciani G, Crivori P, Carrupt PA, Testa B: **Molecular fields in quantitative structure-permeation relationships: The VolSurf approach.** *J Mol Struct* 2000, **503**:17-30.

37. Gund P, Wipke WT, Langridge R: **Computer searching of a molecular structure file for pharmacophoric patterns.** *Comput Chem Res Educ Technol* 1974, **3**:5-21.
38. Beusen DD, Marshall GR: **Pharmacophore definition using the active analog approach.** In *Pharmacophore Perception, Development and Use in Drug Design*. Edited by Güner O. La Jolla: International University Line; 2000:21-45.
39. Güner O (Ed.): *Pharmacophore Perception: Development and Use in Drug Design*. La Jolla: International University Line; 2000.
40. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S: **GRID-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors.** *J Med Chem* 2000, **43**:3233-3243.
41. Johnson MA, Maggiora GM: *Concepts and Applications of Molecular Similarity*. New York: Wiley; 1990.
42. Tanimoto TT: **Non-linear model for a computer assisted medical diagnostic procedure.** *Trans NY Acad Sci Ser 2* 1961, **23**:576-580.
43. Tversky A: **Features of similarity.** *Psychol Rev* 1977, **84**:327-352.
44. Willett P: *Similarity and Clustering Techniques in Chemical Information Systems*. Letchworth: Research Studies Press; 1987.
45. Willett P: **Chemoinformatics – similarity and diversity in chemical libraries.** *Curr Opin Biotechnol* 2000, **11**:85-88.
46. Lewis RA, Pickett SD, Clark DE: **Computer-aided molecular diversity analysis and combinatorial library design.** *Rev Comput Chem* 2000, **16**:1-51.
47. Martin YC: **Diverse viewpoints on computational aspects of molecular diversity.** *J Comb Chem* 2001, **3**:231-250.
48. Oprea TI: **Chemical space navigation in lead discovery.** *Curr Opin Chem Biol* 2002, **6**:384-389.
49. Todeschini R, Consonni V: *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH; 2000.
50. Grant JA, Gallard MA, Pickup BT: **A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape.** *J Comput Chem* 1996, **17**:1653-1666.
51. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P: **Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking.** *J Med Chem* 2003, **46**:2656-2662.
- The methodology and results are clearly documented.
52. Wang R, Liu L, Lai L, Tang Y: **SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex.** *J Mol Model* 1998, **4**:379-394.
53. Peng H, Huang N, Qi J, Xie P, Xu C, Wang J, Yang C: **Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening.** *Bioorg Med Chem Lett* 2003, **13**:3693-3699.
54. Grüneberg S, Stubbs MT, Klebe G: **Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation.** *J Med Chem* 2002, **45**:3588-3602.
- Impressive use of several consecutive filters in the VS process.
55. Brenk RL, Naerum L, Graedler U, Gerber H-D, Garcia GA, Reuter K, Stubbs MT, Klebe G: **Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis.** *J Med Chem* 2003, **46**:1133-1143.
56. Schapira M, Raaka BM, Das S, Fan L, Totrov M, Zhou Z, Wilson SR, Abagyan R, Samuels HH: **Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking.** *Proc Natl Acad Sci USA* 2003, **100**:7354-7359.
57. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 1997, **46**:3-26.
58. Alvarez J: **High-throughput docking as a source of novel drug leads.** *Curr Opin Chem Biol* 2004, **8**:this issue.
59. Schneider G, Neidhart W, Giller T, Schmidt G: **"Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening.** *Angew Chem Int Ed Engl* 1999, **38**:2894-2896.
60. Naerum L, Nørskov-Lauritsen L, Olesen PH: **"Scaffold hopping" and optimization towards libraries of glycogen synthase kinase-3 inhibitors.** *Bioorg Med Chem Lett* 2002, **12**:1525-1528.
61. Singh J, Vlijmen Hv, Liao Y, Lee W-C, Cornebise M, Harris M, Shu I, Gill A, Cuervo JH, Abraham WM, Adams SP: **Identification of potent and novel  $\alpha 4\beta 1$  antagonists using *in silico* screening.** *J Med Chem* 2002, **45**:2988-2993.
- The authors use all available information before conducting the VS experiment.
62. Greene J, Kahn S, Savoj H, Sprague P, Teig S: **Chemical function queries for 3D database search.** *J Chem Inf Comput Sci* 1994, **34**:1297-1308.
63. Flohr S, Kurz M, Kostenis E, Brkovich A, Fournier A, Klabunde T: **Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II.** *J Med Chem* 2002, **45**:1799-1805.
- The use of NMR information, the methodology and results are well documented.
64. Jorgensen WL: **The many roles of computation in drug discovery.** *Science* 2004, **303**:1813-1818.
65. Davis AM, Riley RJ: **Predictive ADMET studies. The challenges and the opportunities.** *Curr Opin Chem Biol* 2004, **8**:this issue.
66. Williams PA, Cosme J, Ward A, Angove HC, Matak-Vinković D, Jhoti H: **Crystal structure of human cytochrome P450 2C9 with bound warfarin.** *Nature* 2003, **424**:464-468.
67. Zamora I, Afzelius L, Cruciani G: **Predicting drug metabolism: A site of metabolism prediction tool applied to the cytochrome P450 2C9.** *J Med Chem* 2003, **46**:2313-2324.
68. Afzelius L, Zamora I, Masimirembwa CM, Karlen A, Andersson TB, Mecucci S, Baroni M, Cruciani G: **Conformer- and alignment-independent model for predicting structurally diverse competitive CYP2C9 inhibitors.** *J Med Chem* 2004, **47**:907-914.
- This method provides a solution to the alignment/conformation choice – a known shortcoming in the 3D-QSAR technology.
69. Boyer S, Zamora I: **New methods in predictive metabolism.** *J Comput Aided Mol Des* 2002, **16**:403-413.
70. Cruciani G, Pastor M, Clementi S, Clementi S: **GRIND (GRID independent descriptors) in 3D structure-metabolism relationships.** In *Rational Approaches to Drug Design*. Edited by Hölting HD, Sippl W. Barcelona: Prous Science Press, 2001:251-260.
71. Cruciani G, Meniconi M, Carosati E, Zamora I, Mannhold R: **VOLSURF: a tool for drug ADME-properties prediction.** In *Drug Bioavailability, Methods and Principles in Medicinal Chemistry, Vol. 18*. Edited by Van de Waterbeemd H, Lennernäs H, Artursson P. Weinheim: Wiley-VCH; 2003: 406-419.
- The effectiveness of VolSurf for ADMET prediction is demonstrated on a wide array of properties.
72. Crivori P, Cruciani G, Carrupt PA, Testa B: **Predicting blood-brain barrier permeation from three-dimensional molecular structure.** *J Med Chem* 2000, **43**:2204-2216.
73. Artursson P: **Epithelial transport of drugs in cell culture. I: a model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells.** *J Pharm Sci* 1990, **79**:476-482.
74. Norinder U, Haeberlein M: **Calculated molecular properties and multivariate statistical analysis in absorption prediction.** In *Drug Bioavailability, Methods and Principles in Medicinal Chemistry, Vol. 18*. Edited by Van de Waterbeemd H, Lennernäs H, Artursson P. Weinheim: Wiley-VCH; 2003:358-405.
75. Abraham MH, Le J: **The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship.** *J Pharm Sci* 1999, **88**:868-880.
76. Huuskonen J: **Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology.** *J Chem Inf Comput Sci* 2000, **40**:773-777.

77. Meniconi M: *Solubility for Potential Drugs, Theoretical and Experimental Methods*. Laurea Thesis (MSc), Perugia: University of Perugia, Italy; 2000.
78. Kratochwil NA, Huber W, Muller F, Kansy M, Gerber PR: **Predicting plasma protein binding of drugs: a new approach**. *Biochem Pharmacol* 2002, **64**:1355-1374.  
This paper includes a comprehensive analysis of literature data and provides additional experimental results.
79. Lombardo F, Obach R, Scott R, Shalaeva MY, Gao F: **Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data**. *J Med Chem* 2002, **45**:2867-2876.
80. Lombardo F, Obach R, Scott R, Shalaeva MY, Gao F: **Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics**. *J Med Chem* 2004, **47**:1242-1250.
81. Vandenberg JJ, Walker BD, Campbell TJ: **HERG K<sup>+</sup> channels: Friend or foe**. *Trends Pharmacol Sci* 2001, **22**:240-246.
82. Pearlstein R, Vaz R, Rampe D: **Understanding the structure-activity relationship of the human ether-a-go-go-related gene cardiac K<sup>+</sup> channel. A model for bad behavior**. *J Med Chem* 2003, **46**:2017-2022.  
Excellent overview of the structure-activity relationships for hERG.
83. Pearlstein R, Vaz R, Kang J, Chen XL, Preobrazhenskaya M, Shchekotikhin AE, Korolev AM, Lysenkova LN, Miroshnikova OV, Hendrix J, Rampe D: **Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches**. *Bioorg Med Chem Lett* 2003, **13**:1829-1835.
84. Cavalli A, Poluzzi E, De Ponti F, Recanatini M: **Toward a pharmacophore for drugs inducing the long QT syndrome: Insights from a CoMFA study of HERG K<sup>+</sup> channel blockers**. *J Med Chem* 2002, **45**:3844-3853.
85. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G: **A virtual screening method for prediction of the hERG potassium channel liability of compound libraries**. *ChemBioChem* 2002, **3**:455-459.  
Simple and effective modeling for hERG binding is given from a large drug dataset.
86. Bleicher KH, Böhm HJ, Müller K, Alanine AI: **A guide to drug discovery: Hit and lead generation: beyond high-throughput screening**. *Nat Rev Drug Discov* 2003, **2**:369-378.
87. Oprea TI: **Property distribution of drug-related chemical databases**. *J Comput Aided Mol Des* 2000, **14**:251-264.
88. Lyne PD: **Structure-based virtual screening: an overview**. *Drug Discov Today* 2002, **7**:1047-1055.
89. Van Drie J: **Pharmacophore discovery: A critical review**.  
• In *Computational Medicinal Chemistry and Drug Discovery*. Edited by Tollenaere J, De Winter H, Langenaeker W, Bultinck P: New York: Marcel Dekker; 2004:437-460.  
Provides a personal account of two decades of pharmacophore discovery.
90. McGann M: **FRED and the future of docking**. OpenEye CUP V, Santa Fe, February 2004. ([http://www.eyesopen.com/about/events/cup\\_v/mcgann/FRED\\_cup5\\_Strategy2.htm](http://www.eyesopen.com/about/events/cup_v/mcgann/FRED_cup5_Strategy2.htm).)
91. Oprea TI, Baroni M, Zamora I, Cruciani G: **High-throughput prediction of passive ADME properties from fragments**. 224th ACS Natl Meeting, Boston, MA, 2002:COMP-109. ([http://www.moldiscovery.com/soft\\_penguins.php](http://www.moldiscovery.com/soft_penguins.php))