Current trends in Virtual High-Throughput Screening using Ligand-based and Structure-based methods

N. Sukumar
Rensselaer Exploratory Center for Cheminformatics Research, Department of Chemistry & Chemical Biology and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

and **Sourav Das**
Department of Chemical Biology and Therapeutics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
Present Address: Proctor & Gamble, India, Ltd., Bangalore, India

Abstract:

High-throughput *in silico* methods have offered the tantalizing potential to drastically accelerate the drug discovery process. Yet despite significant efforts expended by academia, national labs and industry over the years, many of these methods have not lived up to their initial promise of reducing the time and costs associated with the drug discovery enterprise, a process that can typically take over a decade and cost hundreds of millions of dollars from conception to final approval and marketing of a drug. Nevertheless structure-based modeling has become a mainstay of computational biology and medicinal chemistry, helping to leverage our knowledge of the biological target and the chemistry of protein-ligand interactions. While ligand-based methods utilize the chemistry of molecules that are known to bind to the biological target, structure-based drug design methods rely on knowledge of the three-dimensional structure of the target, as obtained through crystallographic, spectroscopic or bioinformatics techniques. Here we review recent developments in the methodology and applications of structure-based and ligand-based methods and target-based chemogenomics in Virtual High-Throughput Screening (VHTS), highlighting some case studies of recent applications, as well as current research in further development of these methods. The limitations of these approaches will also be discussed, to give the reader an indication of what might be expected in years to come.

1. Introduction to Ligand-based and Structure-based VHTS:

The goal of VHTS is to aid in and accelerate the process of design of new drugs and materials with specific desirable physicochemical and/or biological activity profiles. Machine learning, computational pattern recognition or statistical modeling algorithms are employed to generate quantitative correlations between molecular structures and chemical properties or biological activities. The fundamental premise underlying all structure-activity relationship modeling is that molecular structure determines biological activities through the physics of intermolecular interactions. Such modeling can be undertaken either with the knowledge of the structure of the biological target (generally a protein) involved in the activity or even in the absence of any knowledge of the target structure. The former protocol goes by the name of structure-based design. Generally such methods make use of graphic visualization software, shape matching, electrostatic and hydrophobic complementarity for binding site comparisons and for "docking" a small molecule ("ligand") onto the target and parametrized force-field-based scoring functions for estimating the energetics of inter-molecular interactions. In the absence of any detailed knowledge of structure of the target protein, drug design strategies (ligand-based methods) exploit similarities between molecules known to cause the biological activity of interest. The principle here is that changes in molecular structure of the ligand determine changes in biological activity against the same target. Extracting useful information from observed structure-activity relationships in the ligand-based strategy thus requires large-scale data collection, statistical modeling or data mining and a mathematical representation of relevant chemical features of molecules. Statistical modeling techniques can be either classification or regression methods, and can reveal complex relationships between descriptors and biological activity, but it should be borne in mind that such relationships are typically correlative rather than causative — we should be under no illusion that large-scale statistical models enhance our chemical understanding, except in very fortuitous circumstances. The goals of predictive cheminformatics and retrospective model interpretation are, unfortunately, often orthogonal [1].

In the next section we describe strategies and descriptors employed in ligand-based methods, starting with linear models and simple topological descriptors to non-linear models and increasingly complex descriptors, then discuss the important issues of model validation and model applicability domain assessment, *i.e.* how to tell if a model is good enough and to

predict when it will be good. Some currently accepted best practices [2, 3] in predictive cheminformatics are then listed, followed by a discussion of activity cliffs [4-6] or the ruggedness of structure-activity landscapes. Section 3 deals with structure-based methods and section 4 with site similarity approaches and target-based chemogenomics, highlighting some recent work in this rapidly-developing area.

2. Ligand-based VHTS methods:

Linear Free Energy Relationships (LFER)

Descriptors are the intermediary through which molecular structures are represented and key features thereof encoded in a form amenable to computer processing and statistical modeling. Just as there are various ways of representing molecular structure (for example, as one-dimensional alphanumeric SMILES strings, two-dimensional (2D) structure drawings on paper, three-dimensional (3D) molecular models or different cartoon representations popular for macromolecules), there are likewise different families of descriptors that differ in their ease of computation, interpretability, level of detail and in the kinds of molecular features they capture. For a set of molecular descriptors to be useful in modeling, it needs to encode those molecular characteristics pertinent to the property being modeled. When the property of interest is a biological activity, as in drug design, the features to be represented include interactions with other biomolecules that give rise to the biological response. Appropriate representation of inter-molecular interactions is thus critical to successful application of a Quantitative Structure Activity Relationship (QSAR) model. Descriptor utility can be characterized in terms of interpretability and predictive ability on molecules not included in the training set. As indicated by the preceding discussion, these are often mutually conflicting goals. The optimal choice of the descriptor set and the modeling algorithm is often determined by which of these two objectives is more crucial for the problem at hand. Small data sets are best modeled using a small number of well-designed, interpretable descriptors, but this is no guarantee of the performance of the model on new data. The simplest such models are the linear models of classical QSAR that are based on linear free energy relationships (LFER). Hammett [7, 8] first introduced numerical descriptors over 75 years ago to represent the effect of substituent R groups in a molecular scaffold on specified acid ionization equilibria:

$$\log K/K_0 = \sigma\rho \quad (1)$$

where K is the equilibrium constant for the specified reaction, K_0 the equilibrium constant for the reference reaction with R=H, σ is a substituent constant that depends only on R, and ρ is a reaction constant characteristic of the given reaction. Taft [9] expanded the original equation to account for steric effects by introducing a steric substituent constant Es within the linear framework:

$$\log K/K_0 = \rho\sigma + \delta Es \quad (2)$$

These relations have been extensively applied over the years to predict pKa, toxicity and other physicochemical properties. Corwin Hansch [10, 11] extended the LFER concept to describe biological effects of molecules, thereby giving birth to the field of QSAR:

$$\log(1/C) = -k\pi + k'\pi^2 + \sigma\rho + k'' \quad (3)$$

Here C is the concentration of a drug needed to achieve a desired biological activity, π is a lipophilicity parameter introduced by Hansch, and k, k', ρ, and k'' are regression coefficients that were fitted to the training data. Hansch studied the effects of substituents on the partitioning of molecules between two phases: water and octan-1-ol, to model the membrane-aqueous interface in biological systems. The partitioning of a molecule between two different solvents phases is quantified by the equilibrium constant P (partition coefficient) and constitutes an important descriptor of biological activity. The parameter π measures the free energy change caused by a given substituent and is obtained as the difference between the substituted and unsubstituted log P values (π = 0 for H):

$$\pi_R = \log P_R - \log P_H \quad (4)$$

Topological Indices

The concept of descriptors as mathematical characterizations of molecular structure was first introduced by Lamont Kier and coworkers [12-14]. Among the simplest molecular descriptors are those based on atom counts and linear (1D) sequences. Perhaps the most well-known are the descriptors defining Lipinski's Rule of Five [15], which are widely used as a first step in drug design to filter out virtual screening leads with poor bioavailability. Topological

descriptors, which depend only upon the molecular graph or bond connectivity (2D structure), have a long history and, being simple to compute, have proven very useful in QSAR and high-throughput screens [16]. Examples include atom and bond counts, the degree of branching, the number of electrons, the Wiener number W — which counts the lengths of all distances between each pair of atoms in a molecule, Hosoya's topological index Z — which counts all sets of non-adjacent bonds in a structure, and the molecular connectivity index χ, constructed from the row sums of the adjacency matrix (Fig. 1).

Substructural Descriptors

Fragment descriptors are representations of local atomic environments. When the description of the local environment becomes specific, these descriptors are referred to as fingerprints, and are used for substructure searching and for molecular similarity analysis. Binary fingerprints based on 2D structure typically encode the presence or absence of substructural fragments, each describing a substructure of less than ten heavy atoms, a common example being the MACSS key descriptors [17]. Fast searching is accomplished by storing the presence or absence of these fragments as a vector of binary indices, allowing for rapid comparison of molecules in VHTS. Fingerprints based on hashed keys, constructed from atom types, augmented atoms and atom paths, are also popular and implemented in various commercial software programs [18]. Molecular holograms, such as those in Tripos's *Sybyl*® [19], extend keyed fingerprints by storing the frequency of appearance of features, rather than simply their presence or absence. Fingerprints enable rapid similarity searches of large databases, but are not useful for modeling continuous responses. Extended Connectivity Fingerprints (ECFP), developed by SciTegic®, are circular substructural fingerprints [20] where each feature represents an exact structure with limited and specified attachment points, iteratively incorporating information from nearest-neighbors, next-nearest neighbors, *etc*. The set of all features for a neighborhood of specified granularity constitutes a fingerprint; such descriptors can describe both global and local features.

Jürgen Bajorath and coworkers [21] have recently discussed the applicability of 2D fingerprint-based similarity search in scaffold hopping, *i.e.*, the ability to move from one type of structural scaffold or bond framework to a very different one with similar activity. They observed some enrichment for almost all fingerprints evaluated, at approximately the top 1%

of the ranked database, with no single threshold value being applicable to all 2D fingerprint based-similarity search methods for identifying ligands of similar activity. 2D fingerprints capture structural information in a way that makes them difficult to use in scaffold hopping. Structurally diverse compounds typically do not appear in the top ranks in 2D fingerprint based-similarity searches. An approach based on a reduced graph representation can yield more diverse chemotypes than traditional 2D methods [22].

Pharmacophores [23] are a popular type of 3D descriptor, representing specific geometric arrangements of various combinations of atoms of different classes, such as aromatic, lipophilic, positively charged, negatively charged, hydrogen bond donor/acceptor, *etc*. Pharmacophores are designed to represent molecular frameworks that capture the essential geometric features responsible for a drug's biological activity. 3D molecular fingerprints are commonly based on pharmacophore representations of molecules. However, the selection of conformations used to generate the pharmacophores is of crucial importance. If the conformations used to compute descriptors are inappropriate [24], the descriptors contribute primarily noise to the model, and such inappropriate use of 3D descriptors can make a model perform worse than expected (or even worse than simpler 2D descriptors [25]).

Field based models

Another class of 3D descriptor is constructed by encoding conformational information through alignment of molecules in an interaction field. Since biological activity depends upon molecular shape and upon interactions that are primarily non-covalent in nature, molecular mechanics force fields that treat non-covalent interactions as steric and electrostatic forces, are often sufficient to model a large number of biomolecular properties. In the Comparative Molecular Field Analysis (CoMFA®) technique [26], the molecules to be compared are aligned in 3D space by generating the superimposition that maximize the steric and electrostatic overlap or through a pharmacophore model. Once a suitable alignment is obtained, atomic point charges are then calculated for each molecule at a desired level of theory. The next step is to construct a field, for which a probe atom or a group of atoms is chosen to compute steric and electrostatic fields for each molecule at a series of grid points surrounding the aligned data set of molecules in 3D space. The values of the steric and electrostatic fields at each grid point are then used to construct a 3D-QSAR equation, employing a set of molecules with

measured activity as the training set. The predictive quality of a CoMFA 3D-QSAR model largely depends upon the quality of alignment and its resemblance to the actual bio-active conformation. CoMFA employs isocontour plots to represent electrostatically and sterically favorable or unfavorable areas around the molecules, for chemical interpretation of the models. Invariant 3D representations of molecular fragments, called topomers, may also be generated from the corresponding 2D topologies, using deterministic rules that specify absolute configuration, conformation and orientation [27, 28]. Topomers are characterized by CoMFA-like steric shapes and pharmacophore features, and have been employed for shape similarity searching of very large virtual libraries and for generating CoMFA alignments. The steric similarity of topomers is computed as the squared sum of differences in the values of the steric fields at corresponding pairs of lattice points. In contrast to pharmacophore-based 3D searching, where shape comparison focuses on a small set of atom-like features, topomer shape comparison considers all atoms and is computed as a combination of fragment-to-fragment differences, involving a single conformation for each fragment. The main bottleneck in CoMFA is performing the computationally intensive 3D alignments between molecules.

Field-based methods that capture 3D similarity, such as topomers, have the potential to overcome the limitation of 2D fingerprint-based methods with respect to scaffold hopping among different classes of structures [27, 28]. 3D field-based methods were also found to be more effective in scaffold enrichment [29] when diverse compounds were present in the screening database, even when no significant overall performance advantage was observed over 2D fingerprint-based methods. McGaughey, *et al*. [30] observed superior performance of 3D ligand-based methods over docking, but as is usually the case, the performance was dependent upon the data set used for benchmarking.

Local Surface Area Descriptors

Quantifying the relative surface areas of polar atoms (such as N and O) and atomic fragments in molecules, or the surface areas accessible to solvent molecules, is an attractive way to incorporate electrostatic and desolvation effects into 3D QSAR models. Polar surface area (PSA) descriptors [31] obtained from a single conformation often yield predictive models similar to those obtained from averaging multiple conformations [32, 33]. They have been applied in QSAR models for a wide range of biological properties (such as blood-brain

partitioning [33-35], intestinal absorption [32-37] and oral bio-availability [38]). Solvation effects are important in protein folding and stability, while desolvation is associated with protein-protein and protein-ligand binding. The solvent accessible surface area (ASA) [39] measures the solvation free energy of a solute as a sum of atomic contributions, weighted by their solvent-exposed areas. Descriptors based on such local molecular surface properties that do not encode the chemical constitution of a molecule directly, are also likely [40, 41] to favor scaffold hopping and lead to more global QSAR/QSPR models.

Quantum chemistry is, of course, known to be computationally intensive, especially the *ab initio* computations required to generate reliable molecular wave functions and electron densities. Electron density-derived descriptors have thus not been routinely used in VHTS. The Transferable Atom Equivalent (TAE) RECON method [42, 43] overcomes this computational bottleneck by employing a library of atomic charge density fragments and exploiting the theory of Atoms-In-Molecules [44, 45] for the rapid computation of molecular electronic properties from the atomic fragments. Atomic fragments satisfying Bader's virial partitioning prescription [44] have well-defined properties that are approximately additive and transferable from one molecule to another. Molecular descriptors can then be constructed in most cases by simple arithmetic operations on the respective atomic descriptors. For instance, relative surface areas of polar atoms or atoms with electrostatic potential (EP) within a specified range can simply be summed to give the respective molecular surface area descriptor. The atomic density fragments are pre-computed from *ab initio* wave functions and stored, along with the respective atomic descriptors, in a TAE library. At run-time, the RECON algorithm simply matches the atom types in each molecule to the best match from the TAE library and rapidly computes the molecular descriptors from the atomic ones. The algorithm is thus well adapted for virtual high throughput screening applications and scales well with both molecular and database size; throughputs are on the order of a million molecules an hour on a single processor linux computer. Besides the electrostatic potential [46, 47], other commonly used electronic TAE descriptors include Politzer's local average Ionization Potential (PIP) [48, 49], Fukui Reactivity Indices [50, 51] and the Laplacian distribution of the electron density [44], each mapped onto the molecular Van der Waals surface. The local average ionization potential identifies regions with the most tightly-bound (maxima of PIP) and the most ionizable electrons (minima of PIP), the Fukui electrophilic, nucleophilic and radical reactivity indices identify regions most susceptible to electrophilic, nucleophilic and radical

attack respectively, while the Laplacian of the density locates of regions of electron density accumulation or depletion.

A computationally inexpensive set of descriptors that incorporates 3D shape information within the TAE RECON formalism is achieved through property autocorrelation functions, binned by the distance R_{xy} between atom pairs x, y for each TAE property P:

$$A(R_{xy}) = (1/n)\sum_{x,y} P_x P_y \quad (6)$$

Autocorrelation descriptors measure the correlation of a property with itself, measured along the bond path (topological autocorrelations) or through 3D space (spatial autocorrelations). Topological autocorrelation descriptors derived from TAE have been employed [52] to generate improved predictive regression models for the binding affinities of polypeptide sequences to the major histocompability complex.

Shape Descriptors

There are several fast shape comparison methods for virtual high-throughput screening. Rapid Overlay of Chemical Structures (ROCS) [53] is based on the idea that molecules have similar shape if their volumes have significant overlap; any volume mismatch is then a measure of dissimilarity. In ROCS, molecular shape is represented as a continuous function constructed from atom-centered Gaussian functions. The shape of a query molecule is used for scaffold hopping to other molecules with similar 3D shapes, but that may have low similarities to the query molecule in terms of their 2D scaffolds. GRIND descriptors [54-56] represent a class of alignment-independent shape features that encode molecular interaction field distributions at key points around a molecule in the form of correlograms, rather than capturing molecular shape or surface information.

Another fast shape comparison method that avoids the alignment problem is Zauhar's Shape Signature [57, 58]. This method involves a computational ray-tracing procedure within the interior of the molecular envelope (defined by either the van der Waals or solvent-accessible surface) and collecting ray-length and angle-of-reflection information at each point of intersection of the ray with the surface. Shape Signatures encode molecular shape through the distribution of ray lengths (Fig.2), thereby rapidly generating distinctive, compact “shape

signature" fingerprints for each molecule without the necessity of computationally intensive 3D molecular alignments. The Property Encoded Surface Translator (PEST) [59] method extends the concept of Shape Signatures by recording TAE surface property information at each point of intersection of the ray with the molecular surface to produce hybrid shape-property descriptors that are useful for both similarity assessment and constructing QSAR/QSPR regression models. Surface properties that have been fruitfully employed in PEST include electrostatic potentials from either *ab initio* computation or empirical charges, other TAE electronic properties and molecular lipophilicity potentials. Descriptors are encoded as two-dimensional histograms (Fig.3) and wavelet coefficients [60].

Ultrafast shape recognition (USR) was developed as a similarity search tool by Ballester and Richards [61-63]. USR descriptors are molecular shape moments with respect to a small set of well-defined points within a molecule: such as the centroid (*ctd*), the closest atom to the *ctd* (*cst*), the farthest atom from the *ctd* (*fct*) and the farthest atom from the *fct* (*ftf*). Like Shape Signatures or PEST, USR is alignment-free, generates a compact shape profile and has been shown to perform well at shape classification. Furthermore, USR is extremely fast; it is faster than ROCS or even Shape Signatures by several orders of magnitude, and thus well suited for VHTS screens. The RECON [43] algorithm further generates rapid shape-electronic-property hybrid descriptors for high-throughput screening by computing TAE property moments with respect to *ctd*, *cst*, *fct* and *ftf*:

$$USP_{mk} = \sum_i P_i R_{ik}^m \tag{8}$$

for each TAE property P, where the summation i runs over all atoms in the molecule, m = 1, 2, 3 (corresponding to first, second and third moments), and k $\in$ ctd, cst, fct, ftf. While purely shape-based approaches may neglect key aspects of molecular recognition, combining electrostatic information (partial charges or electrostatic potentials) with shape recognition methods (as in RECON [43], PEST [59] and ElectroShape [64, 65]) include electrostatic complementarity in the description. Hybrid descriptors have also be constructed by combining USR with descriptors encoding topological information, such as MACCS keys [66], demonstrating superior recall (ratio of correctly predicted active molecules to the total number of actives) and precision (ratio of correctly predicted active molecules to the total number predicted to be active) in classification models.

Alignment-free descriptors such as USR and Shape Signatures are incapable of chiral discrimination. However, nine of the top ten drugs on the market today have chiral active ingredients. For example, while both enantiomers of Warfarin are anticoagulants, in the case of the widely prescribed beta-blocker Propanolol, the S(-)-enantiomer is approximately 100 times as potent as the R(+)-enantiomer in blocking beta-adrenergic receptors; for another chiral beta-blocker, Sotalol, the enantiomers have equal Class III antiarrhythmic activity, but beta-blocking activity is been attributed mainly to the R-enantiomer; the right-handed S-(+)-form of the painkiller ibuprofen is three times stronger than the left; L-Dopa is used to treat Parkinson's disease, while R-Dopa is toxic! In an effort to meet this challenge of adequately describing biological interactions dependant on chirality, the Richards Group developed Chiral Shape Recognition (CSR) [64, 65], a novel method to compute molecular similarity that builds on the USR method, but distinguishes enantiomers. CSR includes moments with respect to a fourth centroid, defined through a cross product operation. Because of the properties of the cross product, parity inversion changes the signs of all coordinates except that of the fourth centroid, so that the moments with respect to this fourth centroid are different for any molecule and its enantiomer. CSR is of roughly the same speed as USR, since CSR only requires the computationally trivial extra step of computation and renormalization of a single cross-product. Adding chirality and electrostatic complementarity to USR has been shown to result in significant enrichment in virtial screens [64, 65].

Data Fusion

Data fusion was first introduced in the radar sensing community and refers to the process of combining multi-sensor data from different sources, such that the resulting information or model is better than would be possible when these sources are used individually. Data Fusion processes are often categorized under three stages or levels: data level fusion, feature level fusion and decision level fusion. Data level fusion combines several sources of raw data to produce new raw data that is expected to be more informative and synthetic than the inputs. For molecular modeling this is equivalent to combining different sets of descriptors. Several applications along these lines have already been discussed above. Feature level fusion combines various features, as for an example in the combination of several latent variable sets extracted from principal component analysis, partial-least squares analysis, or independent component analysis. Decision fusion combines decisions from several individual

models with either the same or different descriptor sets. Consensus models are example of the latter, where individual model predictions based on the same or different descriptor sets are combined into a single meta-learning model.

Several publications have explored the use of data fusion in molecular modeling and molecular property analysis, for instance by merging similarity scores with molecular descriptors [67-73]. In a recent application, kernel partial-least squares (K-PLS) models with data fusion have shown a significant boost in performance compared to traditional K-PLS models in predicting the binding affinity for the human serum albumin [74]. Several consensus scoring approaches combining highly diverse descriptor sets such as structural keys, property-based fingerprints, shape scores and 3D pharmacophores, have been investigated [75] and shown to give better and more consistent rankings of active molecules. We will encounter further examples of decision-level data fusion when we discuss consensus docking in section 3.

Non-linear Models

The introduction of additional descriptors and the use of non-linear models can add considerable flexibility. However, a proliferation of descriptors or the use of complex, non-linear models often leads to over-fitting (Fig.4) – also known as the “curse of dimensionality”, leading to a tendency for high-capacity models to memorize the details of the training data, thus resulting in a poor ability of the model to generalize the results to data not encountered during training. Over-trained models might give low prediction errors on the training set data, but high errors on test data. Artificial neural networks (Fig.5) are computational models patterned after the learning models prevalent in biological synaptic circuitry, where the descriptors are represented by a set of input neurons and the property to be modeled by an output neuron, connected through one or more “hidden” layers of intervening neurons. The network is first presented with the molecules constituting the training set. During this learning phase, the neural network adjusts the strengths of the connections between the neurons so as to obtain the best match between the output neuron's value (“predicted” response) and the actual measured value of the biological activity being modeled. After training is completed, in the prediction phase, the trained network is presented with the structures for which the response is to be predicted. Neural networks are especially prone to over-fitting, and also suffer from a lack of easy interpretablity of the models. Feature selection methods are

employed to constrain the number of descriptors actually used in the model (and thus the number of adjustable parameters). One such technique is sensitivity analysis [76, 77]: each value of each descriptor is varied within the range spanned by its minimum and maximum values, while holding all the other descriptors frozen at their respective average values, and the change in the predicted response monitored. The descriptors for which the predictions do not vary a lot when they are tweaked are considered less important, and they are gradually pruned from the model in a series of successive iterations between model building and feature selection. Another general feature selection strategy is to add a random variable to the model and to eliminate any descriptor that has a lower correlation with the modeled response than the random variable. Among successful strategies to prevent over-training in neural network models is "early stopping." The data is first split into a training set, an internal validation set and an external test set (Fig.6). Training is stopped before the prediction error on the internal validation set starts to deteriorate. In assessing model performance, it is furthermore important that the model not see the data in the external test set during the training phase.

Statistical modeling techniques that employ capacity control, such as Support Vector Machines (SVM) [78], seek to minimize the sum of the training error and the model complexity (known as the generalization error) instead of minimizing the training error itself, leading to more robust models with better predictive ability on molecules not included in the training set. SVM can be used for classification as well as for regression models: in the former scenario (Fig.7), the goal is to maximize the "margin" ε between the hyperplanes or "support vectors" separating the data points belonging to the different classes (*e.g.* active molecules from inactive ones); in the latter scenario, *i.e.* when employed for a regression between descriptors x_i and a biological activity y (Fig.8), the goal is to minimize the width of the ε-tube between the support vectors within which the "good" data points must fall:

$$\min._{w, b, \xi i, \xi i^*} \{C \sum_i (\xi_i + \xi_i^*) + \tfrac{1}{2} ||\mathbf{w}||^2\} \quad (5)$$

The parameter C controls the tradeoff between training error and capacity, while minimizing $||\mathbf{w}||$ controls the capacity of the linear function:

$$y = (\mathbf{w}.\mathbf{x} + b) \pm \varepsilon. \quad (6)$$

The ε-insensitive loss function applies a steadily increasing penalty to all data points outside the ε-tube, while ensuring that the model is not penalized for not fitting (over-fitting) data points within the ε-tube. SVM belongs a class of statistical methods known as kernel methods, and can be employed with either a linear or a non-linear kernel. A linear kernel is a matrix of linear similarity measures between molecules. A linear SVM with 1-norm regularization ||**w**|| inherently performs feature selection as a side-effect of minimizing capacity in the SVM model, driving many w_i to zero. The basic idea behind feature selection using SVM is very simple: One first constructs a series of sparse linear SVM exhibiting good generalization and finds the subset of variables having nonzero weights in the linear models. This subset of variables is then used in nonlinear SVM to produce the final regression or classification function. A non-linear kernel can be considered as a non-linear data transformation or mapping to a higher-dimensional space, wherein the relationship between the descriptors x_i and the biological activity y can be represented by a linear function. Many different choices for the non-linear kernel are possible; perhaps the most popular is the Radial Basis Function (RBF) or Gaussian kernel:

$$K(x_i,x_j) = \exp\{-\alpha||x_i-x_j||^2\} \qquad \text{for } \alpha > 0 \qquad (7)$$

The data kernel $K(x_i,x_j)$ expresses a non-linear similarity measure between the data.

Model Validation and Applicability Domain

Many authors [79-81, 5, 6] have pointed out that models which perform best, retrospectively, are often the worst prospectively; this observation has been termed the “Kubinyi paradox” . Prospective QSAR on large data sets requires properly validated models. A good cross-validation is no longer sufficient for a model to be considered useful for prospective QSAR [82]. Tests are also required to check for chance correlations; methods include sensitivity analysis described above, where all descriptors having less correlation than an introduced random variable with the activity to be modeled are dropped, and the Y-scrambling test [2], where the activities to be modeled (Y variable) are randomly permuted among the molecules of the training set and any model predicting these scrambled activities well (comparable to the models built with real activities of the training set) is considered suspect and should be discarded. The Y-scrambling test is of particular importance if the data set is small or if the

response variable is categorical (discrete). The scheme outlined in Fig.6 is recommended as a general protocol for obtaining a properly validated, predictive QSAR model.

Methods to characterize not only the predictive ability but also the domain of applicability of models [83-85] are increasingly occupying the attention of researchers. The focus of these investigations is on when it make sense to apply a model within a problem domain. The domain of applicability of a QSAR model is the physicochemical, structural or biological space, the information in which has been used to train the model, and it is within this space that the model is applicable to make predictions for new compounds. The applicability domain of a QSAR model is described in terms of the parameters that are descriptors of the model. Ideally, the QSAR should only be used to make predictions within that domain by interpolation, and not by extrapolation. Compounds which are highly dissimilar from all compounds of the training set (in the space of selected descriptors) can not be predicted with any degree of confidence. Statistical methods to estimate the model applicability domains include [84] range-based (either descriptor range or principal components range may be used), distance-based, geometric and probability-density distribution-based methods. For a given model, two parameters are calculated in the distance-based methods [85]: the average Euclidian distance $<D_k>$ and the standard deviation s_k between each compound of the training set and its k nearest neighbors in the descriptor space. For each test compound i, the distance D_i is calculated as the average of the distances between i and its k nearest neighbors in the training set. The new compound will be predicted by the model only if $D_i \leq <D_k> + Zs_k$, with Z being an empirical parameter. The most straightforward empirical geometric method for defining the coverage of a multi-dimensional set is the convex hull, which is the smallest convex area that contains the original set. The disadvantages of this method are that it can not identify potential empty spaces within the convex hull, and further, the complexity grows as $O(n^{[d/2]+1})$, where n is the number of samples and d the number of dimensions. Probability-density distribution-based methods are the only ones capable of identifying internal empty regions within the convex hull of a QSAR applicability domain.

The use of inappropriate descriptors, injudicious use of high capacity modeling methods without appropriate external validation or without Y-scrambling tests for over-fitting, and application of models outside their demonstrated applicability domain are responsible for most of the problems reported in QSAR modeling in the scientific literature [4, 86]. Other reasons why models fail [81, 2] may include incorrect data (structures and/or activities) in the dataset,

the training set might be too small to model effectively, and activity cliffs (regions of chemistry space where small changes in structure produce huge changes in activity). Before proceeding to a discussion of activity cliffs, we conclude this sub-section by summarizing some of the currently accepted best practices in predictive cheminformatics [2]:

1. There should be a *plausible* (not necessarily a known or a well-understood) mechanism or connection between the descriptors and response. Otherwise we might just as well be doing numerology!
2. Robustness: you cannot keep tweaking parameters until you find one that works just right for a particular problem or data set and then apply it to another. A generalizable model should be applicable across a broad range of parameter space.
3. It is important to know the domain of applicability of the model and stay within it [83-85].
4. Likewise, it is important to know the error bars on the experimental data: there is no point expending a lot of effort modeling the noise in the data.
5. The minimum requirement for developing a predictive model or hypothesis is the "No cheating" principle, *i.e.* no looking at the "answer" or the responses of the prediction set during model building.
6. Divide the data set into training, validation and test sets [82, 2].
7. Validate the training set models using an external validation set [82, 2].
8. Of course, if a data set contains too much noise, no QSAR/QSPR technique can extract a meaningful signal. One should not look too hard for something that may not be there, because one is then liable to be modeling the noise in the data.
9. Consider the use of "filters" to scale and then remove correlated, invariant and "noise" descriptors from the data, and to remove outliers from consideration.
10. Modeling is meant to assist human intelligence – not to replace it. So it is important to try to understand the chemistry of the problem at hand. In this context, however, it is worth re-emphasizing the difference between predictive and retrospective QSAR. Descriptors selected for their ease of interpretation are unlikely to yield optimal predictive models. Conversely, descriptor selection methods designed to generate highly predictive models are often not suitable for mechanistic analysis [1].

Activity cliffs

The fundamental assumption implicit in the VHTS-QSAR protocol, in the context of drug discovery, is that similar molecules should exhibit similar activities in biological assays [87, 88]. While such correlations are often observed for simple physicochemical properties, significant mis-predictions of biological activity still arise among very similar molecules even with the best of validation. Thus for example, Yvonne Martin *et al*. [88] found in a follow-up to 115 high-throughput screening assays that there was only a 30% probability that a compound showing similarity greater than 0.85 to an active molecule would itself be active! Gerry Maggiora [4] postulated that the reason for such deviations might be related to the complex nature of the activity landscape associated with a given biological assay, which in turn is related to the chemical-space representation (molecular descriptor space) used to characterize the set of compounds assayed and to the similarity assessment metric employed. He summarized this with the catch-phrase "Not all chemical spaces are created equal!" Maggiora's topographical metaphor recognizes that very similar molecules may in some cases possess very different activities, giving rise to "activity cliffs" [4, 89], and leading to deviations from the similarity principle. While similar molecules may not always exhibit similar activities in individual biological assays, similar molecules do display similar broad patterns of biological activities across a range of related targets [90, 91], a fact that has been exploited [92, 93] to construct molecular descriptors from activities determined from a broad spectrum of *in vitro* assays representing a cross-section of the druggable proteome. These ideas are exploited further in target-based chemogenomics, which we discuss in section 4.

Guha and Van Drie [5-6] have addressed the identification and the quantification of activity cliffs in chemical models of biological activity by defining a Structure-Activity Landscape Index (SALI):

$$SALI_{i,j} = |A_i - A_j|/\{1 - sim(i,j)\} \quad (8)$$

where A_i and A_j are the activities of the i^{th} and the j^{th} molecules, and sim(i,j) is the similarity coefficient between the two molecules. Steep activity cliffs in a data set lead to high SALI values. These are the most interesting regions of a structure-activity relationship for purposes of drug design. The SALI values for a dataset can be visualized as a heat map, thus identifying and characterizing the cliffs of biological activities. Alternatively, utilizing a cut-off

value of the index enables one to represent pair-wise sets of molecules through network graphs [94, 5-6], thereby highlighting abrupt changes in response associated with the steepest (most significant) cliffs. Guha and Van Drie [5-6] have also defined the SALI curve as a way to assess QSAR models and modeling protocols. The SALI curve is a plot of S(X) versus X, where S(X) is the SALI value at the similarity threshold X. While the SALI network graph orders each pair of molecules by activity, the SALI curve tallies how many of these orderings a model is able to predict.

3. Structure Based Methods

Structure-based methods complement ligand-based approaches when the protein structure is available from crystallographic studies or deduced from homology to known structures. Appropriate structure preparation is a critical step for structure based virtual screening. Some commonly used tools that are part of commercial molecular modeling packages include protonate3D and ProPka in MOE and the protein and ligand preparation workflows available within the Maestro package available from Schrodinger, Inc. The effects of ligand tautomer enrichment on virtual screening results have been the subject of several investigations [95-100]. Miletti and Vulpetti [99] note that conclusions have been drawn both in favor of and against tautomer enrichment. The use of all possible tautomers has been reported to increase computational time [96] and to increase false positives [97, 98]. However, use of different tautomers was not found to have any significant impact in the docking results of Oellien *et al*. [100] on dyhydrofolate reductase, transketolase and R-trichosanthin targets. Miletti and Vulpetti [99] further found that inclusion of the most stable tautomeric form in water had a higher enrichment rate than just including the least stable form, in a docking study using Flap, Glide, and Gold on seven targets of the DUD data set [101]. However, including all forms did not have any significant disadvantage over including only the most stable form in water. Tautomer form reversal was observed in water versus binding site in ligands with low ΔG (<2 kcal mol^{-1}) or those undergoing annular tautomerism.

There are indications that virtual high-throughput screening may complement experimental high-throughput screening [102]. In a study by Babaoglu *et al.* [102, 103], a qHTS campaign involving screening of 70,000 compounds with a maximum concentration of 30 μM against AmpC β-lactamase revealed no true inhibitors. However two false negatives in the qHTS

screen were correctly identified by a docking-based virtual screen that placed these two compounds in ranks 80 and 200 respectively. When experimentally validated by a low throughput method and at a higher compound concentration, K_i of 37 µM and 55 µM respectively were realized. Kolb *et al*. [102] also cite unpublished results involving a qHTS campaign of 198,000 molecules against the enzyme cruzain, which revealed false negatives from a docking based virtual screen.

Kolb *et al.* [102] remarked on the growing interest in using docking as a tool for identification of substrates and agonists and hence function of enzymes [104, 105]. They also pointed to the improved performance of X-ray crystal structures of GPCRs as targets for docking, as opposed to earlier homology models. Where 5% hit rate in docking is considered substantial, the screens that utilized crystal structures of GPCRs in the two cited studies [106, 107] had hit rates of 24% and 36% [102].

Changes in protein structure can affect the performance of docking-based virtual screens; this has been a well-recognized issue with rigid receptor docking. Common benchmark exercises generally involve re-docking into the same cavity from which the ligand was separated. The performance drops significantly when the ligands are cross-docked to the cavity of the same protein but separated from a different bound ligand. Generally programs that make use of a single fixed protein conformation have success rates closer to 20% when applied to cross-docking exercises [108]. New approaches that make use of protein conformations from multiple complexes, such as Surflex-Dock, achieved an average success rate as high as 61% (across eight pharmaceutically relevant targets). Following docking, protein pocket adaptation and rescoring, this program identified single pose families correctly an average of 67% of the time. Considering the best of two pose families from alternate scoring approaches yielded a 75% mean success rate.

Docking and scoring also form an integral part of the fragment-based drug discovery workflow. Chen and Shoichet [109] report that a docking protocol which failed to identify true positive lead-like compounds for the target CTX-M beta-lactamase, nevertheless correctly identified several low-affinity fragment-like molecules of diverse chemotypes as docking hits. This is interesting to note, as fragment-like molecules have greater chance of adapting to a binding pocket in spite of their low flexibility [109]. Furthermore, the predicted binding pose

closely coincided with the observed crystal-structure derived pose. The low-affinity fragment hits were subsequently optimized to inhibitors of higher specificity and higher affinity.

Several successes in fragment-docking have been reported in the recent literature. Mpamhanga *et al*. [110] described the discovery of selective Pteridine Reductase 1 inhibitors by following a fragment-based docking protocol. Two chemical series, aminobenzothiazole and aminobenzimidazole scaffolds, were identified of which one molecule from the latter scaffold was co-crystallized and found to have a binding pose similar to that predicted in the docking study. However, two other analogs of the same series exhibited different binding modes with change in protein conformation and involvement of water molecules. Ekonomiuk *et al.* [111] reported the discovery of a non-peptidic inhibitor of West Nile virus NS3 protease in a fragment-based high-throughput docking campaign, using 22 compounds for experimental validation from an initial library of nearly 12,000 molecules. Six of the 22 compounds showed specific affinity upon validation by NMR spectra of the compounds and the protein. The molecule which showed the most pronounced effects on the NMR spectrum of the protease, was further characterized and shown to bind specifically to the active site with an affinity of about 40 μM.

When comparing ligand-based methods with structure-based methods Krüger and Evers [112] found that ligand-based virtual screening gave enrichment similar to structure-based virtual screening against four targets, namely angiotensin-converting enzyme (ACE), cyclooxygenase 2 (COX-2), thrombin and human immunodeficiency virus 1 (HIV-1) protease. However, the hits were found to be non-overlapping and the authors suggested parallel application of the two approaches. For ligand-based virtual screening, ROCS (3D-similarity searching), Feature Trees and Scitegic Functional Fingerprints (2D-similarity searching) were employed, whereas for structure-based virtual screening a combination of GOLD, Glide, FlexX and Surflex and nine scoring functions were used in the docking programs.

Irwin and coworkers [113] reported the development of an expert docking system and studied the feasibility of unsupervised docking. The system called “Dock Blaster” has six components, each having a specific function from reading a PDB file to preparing reports for screening interpretation. The method was benchmarked against the Astex [114], GOLD [115] and DUD [101] sets and found to have pose fidelity within 2 Å rmsd for about 50−60% of targets, as

compared to 70−80% for expert-guided docking. Both good pose fidelity and good enrichment were observed in 25−40% of benchmark cases. In a further study of 7755 complexes, the re-docked ligand ranked in the top 5% of 100 property-matched decoys in 1398 complexes, with an rmsd less than 2 Å. These results suggest that unsupervised prospective docking could be a viable protocol.

Fan *et al*. [116] studied the applicability of comparative models in structure-based virtual screening and found that a consensus docking approach based on several comparative models from different templates outperformed random selection and was better or comparable overall to the results obtained from the holo and apo X-ray crystal structures. Docking to comparative models was found to be better than docking to the template itself. However, it was not possible to predict which model would have the greatest enrichment.

When performing docking and scoring, the reliability of the scoring method is a cause for concern: the lack of explicit treatment of entropy in a scoring function for virtual high-throughput screening may result in poor affinity estimation whenever entropy plays a dominant role. The converse may also be true, in that such scoring functions may show high accuracy when enthalpy far exceeds entropic contributions to binding. A database of enthalpy and entropy values derived from isothermal titration calorimetry experiments is available from the work of Olsson and coworkers [117]. Analysis of the data set showed the so-called “enthalpy-entropy compensation” [118] that resulted in a relatively small range of variation for free energy and large ranges for entropy and enthalpy. When the enthalpy values were plotted against free-energy values, no correlation between enthalpy and free energy was observed [119] (similar to earlier observations [120]). However, good correlation between free energy and enthalpy was observed for a third of the entries (111 of 332) where the enthalpy was greater than entropy by at least a factor of three. This trend seemed to reflect the trend in accuracy of scoring functions: in benchmarking studies of scoring functions [121, 122], trypsin consistently emerged as an easy target (higher accuracy of scoring functions), whereas scoring functions had poor accuracy in predicting affinities for the HIV protease target. From the available data in SCORPIO, trypsin complexes showed a higher enthalpy contribution on average than HIV protease complexes.

While further improvement in reliability of scoring functions across diverse protein targets is

highly desirable, several success stories have already been documented and many others are expected to follow.

4. Site Similarity Approaches and Target based Chemogenomics

There has been increasing interest in automated high-throughput comparison of protein binding sites [123-126] in recent years, one reason being the shift from traditional receptor-specific studies to a cross-receptor view [90]. Because different proteins may have different sequences or folds and yet have similar binding partners, it is necessary to compare binding sites for establishing relationships between protein targets. Work from Brian Shoichet's group [127] and by Paolini, *et al.* [128] has shown that ligands quite frequently have affinity for more than one target. While similar ligands may bind to similar targets [90], ligand similarity in itself is limited by the accuracy of ligand-similarity detection algorithms. As shown by Kinnings, *et al*. [129] and by Das, *et al.* [130], ligands that would otherwise be considered dissimilar by commonly used ligand-similarity detection algorithms can bind on to the same target. This is where protein target based approaches can fill the gap and identify related targets by comparing their binding sites, thereby pinpointing the potential for cross-reactivity between the corresponding ligands. A target virtual screen can, in addition, rapidly detect potential leads for new targets and offer significant scope for identifying alternate ligand chemotypes of similar activity.

Binding site comparison techniques may utilize the shape of the binding site, the properties of the cavity lining residues or a combination thereof, the relative positions of the atoms or pseudo-centers derived from the atomic coordinates, or simply residue identities for automated comparison. Methods that rely heavily on geometry must ideally be able to accommodate changes occurring at the binding site due to receptor rearrangement upon binding to different ligands of dissimilar shape and/or sizes. Another challenge is to recognize similarity in binding sites that have low sequence identity but bind to the same ligand in significantly different ligand conformations.

In the post-genomic era we are seeing a convergence between the goals and methods of bioinformatics and cheminformatics [131], catalyzed by rapid advances in the field of chemogenomics and by the greater availability of high-throughput data (structure, binding

affinity and functional effects) for both targets and ligands of pharmaceutical interest. Chemogenomics [132, 133] refers to the science of relating the protein target space to the ligand space. Chemogenomic approaches can aid in drug discovery where ligand information is sparse or where structural information is absent [90]. In an excellent review of ligand and target based chemogenomic approaches, Rognan [91] highlighted specific examples and approaches where chemogenomics revealed important insights into the drug discovery process. In this section, we focus on target-based chemogenomics and cite specific examples where binding site comparison has led to new lead discovery or rationalization of reactivity profiles of drugs across protein targets.

Oloff, *et al.* developed a novel structure-based cheminformatics approach, Complementary Ligands Based on Receptor Information (CoLiBRI) [134], to search for ligands complimentary to binding sites. This method was based on the representation of both receptor binding sites and their respective ligands in a space of universal TAE chemical descriptors. Knowledge of the receptor active site structure enabled identification of the known complimentary ligand among the top 1% of a large chemical database in over 90% of all test cases, when a binding site of the same protein family was present in the training set. They also demonstrated identification of complementary receptor sites starting from a ligand chemical structure.

The sequence order independent profile-profile alignment (SOIPPA) developed by Xie and Bourne [135] detects similar binding sites of proteins unrelated in sequence or even function. The algorithm has been implemented in the web server SMAP-WS for proteome-wide ligand-binding site comparison [136]. Using this algorithm, the relationship between the cofactor binding sites of NAD-binding Rossmann folds and the SAM-binding domain of the SAM-dependent methyltransferases was established [129]. Following up of this study, entacapone and tolcapone, which are known to bind to COMT (a SAM-dependent methyltransferase in the presence of the SAM cofactor), were docked into 215 NAD-binding proteins. Among these, the NAD-binding InhA enzyme that is the target for anti-tuberculosis drugs consistently appeared among the top-scoring targets. On alignment of InhA and COMT using the SOIPPA method, the cofactor- and ligand-binding pockets were found to coincide. Entacapone on subsequent experimental validation showed a MIC99 of 260 μM for mycobacterium tuberculosis, well below the *in vitro* toxicity limit determined from a neuroblastoma cell line. Interestingly, the authors found low 2D similarity between the ligands of InhA and entacapone.

In a similar observation by Das *et al*. [119, 130], a geometry-based binding-site comparison method utilizing the Property Encoded Shape Distributions (Figs.9-10) was able to detect high similarity between sites that had cross-reacting ligands of low 2D and 3D similarities. The method is available as a web server [137] for high-throughput screening of a ligand-bound site against a database of over a hundred thousand ligand-bound sites derived from the PDB. The algorithm uses a surface representation of the sites, and has the potential to correctly detect similarities among relevant sites even when sequence identities at the sites are low.

Martin *et al*. [138] described the discovery of selective, nonpeptidic, small-molecule somatostatin receptor subtype 5 (SST5R) antagonists by applying chemogenomics approaches. From the similarity of the putative ligand-binding residues of SST5R and opioid, histamine, dopamine and serotonine receptors, astemizole was identified as a lead. Astemizole is an antagonist for the H1 receptor; the structure was subsequently modified to have nanomolar affinity for SST5R and no activity for H1 receptor. Gloriam *et al*. [139] recently characterized the selectivity profile of ligands at Family A GPCRs by sequence similarity at the ligand binding regions. From the analysis of aminergic binding sites, the authors concluded that it was possible to classify receptors in a way that reflects their binding affinities, if the ligand binding residues were known.

Several studies have tried to rationalize selectivity profiles of kinase inhibitors. Kinnings and Jackson [140] used geometric hashing to identify similarities in binding sites of kinases and to predict binding partners for kinase inhibitors. They also recognized the potential of this approach in protein-based “inverse” virtual screening. Sheridan *et al*. [141] used property descriptors derived from ligand-binding residues to rationalize kinase inhibitor polypharmacology (the interaction of a drug with multiple targets), while more recently Milletti and Vulpetti [142] used a Shape Context [143] based binding site similarity detection approach to rationalize polypharmacology. This growing appreciation for the role of polypharmacology is leading to a paradigm shift from traditional receptor-specific to a cross-receptor view and spurring the application of network approaches [144] to drug discovery and development, with the goal of expanding the opportunity space for new drugs through designing for improved selectivity and efficacy and lower toxicity. Rational drug design using target-based chemogenomics thus complements high-throughput screening for finding better starting points for a drug discovery program.

As more and more structural information on protein targets becomes available, receptor-based virtual screening approaches will continue to gain prominence and complement ligand-based approaches. Commercial molecular modeling suites are already including binding site comparison modules [124] and several free web-based servers are now available. With the concept of a “magic shotgun” replacing the “magic bullet” [145] one can expect rapid developments in the field of target-based chemogenomics in the near future, geared towards addressing challenges arising from receptor and ligand flexibility, and solving the important cross-reactivity problem.

5. Conclusions

VHTS aims to aid and accelerate the drug discovery process through application of machine learning and statistical modeling algorithms, to generate quantitative correlations between molecular structures and biological activities. The fundamental premise underlying structure-activity relationship modeling is that molecular structure determines biological activities. In the absence of detailed knowledge of the structure of the target protein, ligand-based methods are employed, exploiting similarities between molecules known to be biologically active. A variety of molecular descriptors may be employed to represent and encode structures for computerized analysis and modeling. While statistical modeling techniques can reveal complex relationships between descriptors and biological activity, they may not always enhance our chemical understanding, because correlation does not imply causation. Predictive cheminformatics and retrospective model interpretation are often non-overlapping goals. Excessively noisy data or insufficient data for training, inappropriate descriptors, use of high capacity modeling methods without appropriate external validation or tests for over-fitting, the presence of activity cliffs and application of models outside their demonstrated domain of applicability are some of the reasons why models fail. To safeguard against these pitfalls in predictive cheminformatics, it is important to adopt and follow the best practices in the field. Activity cliffs can lead to very similar molecules having very different activities in biological assays. While similar molecules may not always exhibit similar activities in individual assays, they do display similar broad patterns of biological activities across a range of related targets. Chemogenomic approaches, which relate the protein target space to the ligand space, are especially useful where ligand information is sparse. With the steadily growing body of data on protein structures, structure-based models including docking-based

virtual screening, high-throughput protein binding site comparison and target-based chemogenomics are becoming increasingly common, helping to leverage the information available on biological targets and to complement ligand-based approaches. We can thus expect to see continuing advances in rational drug design along these lines.

References:

[1] Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus ranking approach to understanding the underlying mechanism with QSAR. *J. Chem. Inf. Model.* **2010**.

[2] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation, *Molecular Informatics* **2010**, 29 (6-7), 476–488.

[3] Breneman, C. M. Predictive Cheminformatics: Best Practices for Determining Model Domain Applicability, *Sanibel Conference*, Sanibel Island, Florida, 2007 http://reccr.chem.rpi.edu/Presentations/Sanibel2007_BestPractices.pdf

[4] Maggiora, G.M. On outliers and activity cliffs — Why QSAR often disappoints. *J. Chem. Inf. Model.*, **2006**, 46, 1535.

[5] Guha, R; Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs, *J. Chem. Inf. Model.* **2008**, 48, 646-658.

[6] Guha, R; Van Drie, J. H. Assessing how well a modeling protocol captures a structure-activity landscape, *J. Chem. Inf. Model.* **2008**, 48, 1716–1728.

[7] Hammett, L. P. Some relations between reaction rates and equilibrium constants. *Chem. Rev.* **1935,** 17, 125.

[8] Hammett, L. P. The effect of structure upon the reactions of organic compounds. benzene derivatives. *J. Amer. Chem. Soc.* **1937,** 59, 96.

[9] Taft, R. W. Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J. Amer. Chem. Soc.* **1952,** 74, 3120.

[10] Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. *J. Amer. Chem. Soc.* **1963,** 85, 2817-2824.

[11] Hansch, C.; Leo, A., *Exploring QSAR*. ed.; American Chemical Society: 1995.

[12] Kier, L. B.; Hall, L. H., *Molecular Connectivity in Chemistry and Drug Research*. ed.; Acadmic Press: New York, 1976.

[13] Kier, L. B.; Hall, L. H., *Molecular Connectivity in Structure-Activity Analysis*. ed.; Research Studies Press: Letchworth, 1986.

[14] Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular connectivity I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **1975,** 64, 1971-1974.

[15] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001,** 46 (1-3), 3.

[16] Randić, M. The connectivity index 25 years after. *J. Mol. Graph. Model.* **2001,** 20 (1), 19-35.

[17] Elsevier MDL®, San Ramon, CA.

[18] Daylight® Chemical Information Systems, Inc. Aliso Viejo, CA.

[19] *Sybyl®*, Tripos, L.P. St. Louis, MO

[20] Morgan, H. L. *J. Chem. Doc.* **1965,** 5, 107-113.

[21] Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening, *J. Med. Chem.* **2010** *53* (15), 5707-5715

[22] Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs, *J. Chem. Inf. Model.* **2006**, *46* (2), 503-511.

[23] Ehrlich. *Dtsch. Chem. Ges.* **1909,** 42, 17.

[24] Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein−ligand binding. *J. Med. Chem.* **2006,** 49, 5880-5884.

[25] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004,** 44, (1), 170-178.

[26] Cramer, I., Richard, D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteIns. *J. Amer. Chem. Soc.* **1988,** 110, 5959.

[27] Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer similarity searching of conventional structure databases. *J. Molec. Graph. Model.* **2002,** 20, 447-462.

[28] Jilek, R. J.; Cramer, R. D. Topomers: A Validated Protocol for Their Self-Consistent Generation. *J. Chem. Inf. Comput. Sci.* **2004,** 44, 1221-1227.

[29] Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS, *J. Chem. Inf. Model.* **2008** *48* (4), 719-729.

[30] McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. .; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening, *J. Chem. Inf. Model.* **2007** *47* (4), 1504-1519.

[31] Saunders, R. A.; Platts, J. A. Correlation and prediction of critical micelle concentration using polar surface area and LFER methods. *J. Phys. Org. Chem.* **2004,** 17 (5), 431-438.

[32] Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharmaceut. Sci.* **1996,** 85 (1), 32-9.

[33] Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci* **1999,** 88, 807–814.

[34] Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration. *J. Pharm. Sci.* **1999,** 88 (8), 815-821.

[35] Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharmaceut. Res.* **1999,** 16 (10), 1514.

[36] Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmac. Res.* **1997,** 14 (5), 568.

[37] Camenisch, G.; Folkers, G.; van de Waterbeemd, H. *Pharm. Acta Helvet.* **1996,** 71, 309–327.

[38] Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002,** 45 (12), 2615-2623.

[39] Eisenberg, D.; McClachlan, A. Solvation energy in protein folding and binding. *Nature* **1986,** 319, 199-203.

[40] Ehresmann, B.; Groot, M. J. d.; Alex, A.; Clark, T. New molecular descriptors based on local properties at the molecular surface and a boiling-point model derived from them. *J. Chem. Inf. Comput. Sci.* **2004,** 44 (2), 658 -668.

[41] Clark, T. QSAR and QSPR based solely on surface properties? *J. Molec. Graph. Model.* **2004,** 22, 519–525.

[42] Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D. Transferable atom equivalent multicentered multipole expansion method. *J. Comput. Chem.* **2003**, 24(4), 512-529.

[43] Sukumar, N.; Breneman, C. M. QTAIM in drug discovery and protein modeling, in *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*, Matta, C.F.; Boyd, R.J., Eds.; Wiley-VCH, **2007**, pp.471-498.

[44] Bader, R. F. W., *Atoms in Molecules: A Quantum Theory*; Oxford Press: Oxford, **1990**.

[45] Matta, C. F.; Boyd, R. J. *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*; Wiley-VCH: Weinheim, **2007**.

[46] Politzer, P.; Murray, J. S.; Peralta-Inga, Z. Molecular surface electrostatic potentials in relation to noncovalent interactions in biological systems. *Int. J. Quantum Chem.* **2001,** 85 (6), 676-684.

[47] Politzer, P.; Truhlar, D. G., *Chemical Applications of Atomic and Molecular Electrostatic Potential*; Plenum Press: New York, **1981**.

[48] Murray, J. S.; Politzer, P. *Theoret. Computat. Chem.* **1998,** 5, 198-202.

[49] Murray, J. S.; Politzer, P.; Famini, G. R. Theoretical alternatives to linear solvation energy relationships. *J. Molec. Struct. (THEOCHEM)* **1998,** 454 (2-3), 299-306.

[50] Fukui, K. Role of frontier orbitals in chemical reactions. *Science* **1987,** 218, 747-754.

[51] Parr, R. G.; Yang, W. Density functional approach to the frontier-electron theory of chemical reactivity. *J. Amer. Chem. Soc.* **1984,** 106, 4049-4050.

[52] Bergeron, C.; Hepburn, T.; Sundling, M.; Sukumar, N.; Bennett, K. P.; Breneman, C. Prediction of peptide bonding affinity: kernel methods for nonlinear modeling. *Protein Peptide Lett., in press*.

[53] Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-d scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005,** 48, (5), 1489-1495.

[54] Cruciani, G.; Watson, K. Comparative Molecular Field Analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, 37, 2589-2601.

[55] Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting blood−brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, 43, 2204-2216.

[56] 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.

[57] Zauhar, R.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J. Shape Signatures: A new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003,** 46, 5674-5690.

[58] Nagarajan, K.; Zauhar, R.; Welsh, W. J. Enrichment of ligands for the serotonin receptor using the shape signatures approach. *J. Chem. Inf. Model.* **2005,** 45, 49-57.

[59] Breneman, C. M.; Sundling, C. M.; Sukumar, N.; Shen, L.; Katt, W. P.; Embrechts, M. J. New developments in PEST shape/property hybrid descriptors. *J. Comput.-Aided Mol. Des.* **2003***,* 17, 231-240.

[60] Sundling, C. M.; Sukumar, N.; Zhang, H.; Embrechts, M. J.; Breneman, C. M. Wavelets in chemistry and cheminformatics. *Rev. Comput. Chem.* **2006,** 22, 295-329.

[61] Ballester, P. J.; Richards, W. G. Ultrafast shape recognition for similarity search in molecular databases, *Proc. R. Soc. A* **2007**, 463, 1307-1321.

[62] Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, 28, 1711–1723.

[63] Ballester, P. J.; Finn, P. W.; Richards W. G. Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J. Mol. Graph. Model.* **2009**, 27, 836–845.

[64] Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Richards, W. G. Molecular similarity including chirality. *J. Mol. Graph. Model.* **2009**, 28, 368–370.

[65] Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Moretti, L.; Cooper, R. I.; Richards, W. G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput.-Aided Mol. Des.* **2010**, 24, 789–801.

[66] Cannon, E. O.; Nigsch, F.; Mitchell, J. B. O. A novel hybrid ultrafast shape descriptor method for use in virtual screening. *Chem. Central J.* **2008**, 2, 3.

[67] Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1177-1185.

[68] Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and use of fragment-occurrence data in similarity-based virtual screening. *J. Comput.-Aided Mol. Des.* **2009**, 23, 655–668.

[69] Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Persp. Drug Disc. Design*. **2000**, 20.

[70] Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 435-442.

[71] Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**.

[72] Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C., Morin-Allory, L. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 276-285.

[73] Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional probability: a new fusion method for merging disparate virtual screening results. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 601-9.

[74] Huang, C.; Embrechts, M. J.; Sukumar, N.; Breneman, C. M. Data fusion and auto-fusion for quantitative structure-activity relationship (QSAR). *Lecture Notes Comput. Sci.* ICANN (Springer) **2007**, 1, 628-637.

[75] Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.***2006**, 46, 277-288.

[76] Embrechts, M. J.; Arciniegas, F. A.; Ozdemir, M.; Kewley, R. H. (2003) "Data Mining for Molecules with 2-D Neural Network Sensitivity Analysis" *International Journal of Smart Engineering System Design*, 5, 225-239.

[77] Embrechts, M. J.; Bress, R. C.; Kewley, R. H. Feature selection via senitivity analysis with direct kernel PLS" *Feature Selection in Machine Learning*. Guyon, I; Gunn, S., Eds., Springer Verlag, *Stud. Fuzz.* **2006**, 207, 447–462.

[78] Vapnik, V, N. Statistical Learning Theory, Wiley-Interscience, **1966**.

[79] Norinder, U. Single and domain made variable selection in 3D QSAR applications. *J. Chemom.* **1996**, 10, 95–105.

[80] Kubinyi, H.; Hamprecht, F.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, 41, 2553–2564.

[81] Kubinyi, H. Why Models Fail http://www.kubinyi.de/san-francisco-09-06.pdf

[82] A. Golbraikh, A. Tropsha, “Beware of q2 !”, J. Mol. Graph. Model. 2002, 20, 269–276.

[83] Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; O¨berg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection, *J. Chem. Inf. Model.* **2008**, 48 (9), 1733–1746.

[84] Nina Nikolova and Joanna Jaworska, Approaches to Measure Chemical Similarity - a Review, *QSAR Comb. Sci.* **2003**, 22, 1006-1026.

[85] Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest. *QSAR Comb. Sci.* **2003**, 22, 69-77.

[86] Johnson, S. R. The trouble with QSAR (or how i learned to stop worrying and embrace fallacy), *J. Chem. Inf. Model.* **2008**, 48 (1), 25–26.

[87] Johnson, M.; Maggiora, G. *Concepts and Applications of Molecular Similarity*. John Wiley, New York, **1990**.

[88] Martin, Y. C.; Kofron, J.L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, 45, 4350-4358.

[89] Peltason, L.; Bajorath, J. SAR Index: quantifying the nature of structure-activity relationships, *J. Med. Chem.* **2007**, 50, 5571-5578.

[90] Klabunde, T., Chemogenomic approaches to drug discovery: similar receptors bind similar ligands, *Br. J. Pharmacol.* **2007**, 152 (1), 5–7.

[91] Rognan, D., Chemogenomic approaches to rational drug design, *Br. J. Pharmacol.* **2007**, 152, 38–52.

[92] Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra analysis: Model proteome characterization for linking molecular structure and biological response. *J. Med. Chem.* **2005**, 48, 6918-6925.

[93] Fliri, A. F.; Loging, W. T. .; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Nat. Acad. Sci. USA* **2005**,102, 261-266.

[94] Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices, *J. Med. Chem.* **2008**, 51, 6075–6084.

[95] ten Brink T, Exner TE. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J Chem Inf Model.* **2009**, 49 (6), 1535-46.

[96] Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J. Chem. Inf. Model.* **2009**, 49, 2742–2748.

[97] ten Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, 49, 1535–1546.

[98] Clark, R. D.; Shepphird, J. K.; Holliday, J. The effect of structural redundancy in validation sets on virtual screening performance. *J. Chemom.* **2009**, 23, 471–478.

[99] Milletti F, Vulpetti A. Tautomer preference in PDB complexes and its impact on structure-based drug discovery. *J. Chem. Inf. Model.* **2010**, 50, 1062–1074

[100] Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W. D.; Selzer, P. M. The impact of tautomer forms on pharmacophore-based virtual screening. *J. Chem. Inf. Model.* **2006**, 46, 2342–2354.

[101] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, 49, 6789–6801.

[102] Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr Opin Biotechnol.* **2009**, 20 (4), 429-36.

[103] Babaoglu, K.; Simeonov, A.; Irwin, J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C.; Shoichet, B. K. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against b-lactamase. *J. Med. Chem.* **2008**, 51, 2502-2511.

[104] Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. Structure-based activity prediction for an enzyme of unknown function. *Nature* **2007**, 448, 775-779.

[105] Xiang, D. F.; Kolb, P.; Fedorov, A. A.; Meier, M. M.; Fedorov, L. V.; Nguyen, T. T.; Sterner, R.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. Functional annotation and three-dimensional structure of Dr0930 from Deinococcus radiodurans, a close relative of phosphotriesterase in the amidohydrolase superfamily. *Biochemistry* **2009**, 48, 2237-2247.

[106] Kolb, P.; Rosenbaum, D. M.; Irwin, J. J.; Fung, J.; Kobilka, B. K.; Shoichet, B. K. Structure-based discovery of b2-adrenergic receptor ligands. *Proc. Natl. Acad. Sci. USA* **2009**, 106, 6843-6848.

[107] Sabio, M.; Jones, K.; Topiol, S. Use of the X-ray structure of the b2-adrenergic receptor for drug discovery. Part 2. Identification of active compounds. *Bioorg. Med. Chem. Lett.* **2008**, 18, 5391-5395.

[108] Jain, A. N. Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J. Comput.-Aided Mol. Des.* **2009**, 23 (6), 355-74.

[109] Chen, Y.; Shoichet B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2009**, 5(5), 358-64.

[110] Mpamhanga, C. P.; Spinks, D.; Tulloch, L. B.; Shanks, E. J.; Robinson, D. A.; Collie I. T.;

Fairlamb, A. H.; Wyatt, P. G.; Frearson, J. A.; Hunter, W. N.; Gilbert, I. H.; Brenk. R. One scaffold, three binding modes: novel and selective pteridine reductase 1 inhibitors derived from fragment hits discovered by virtual screening. *J. Med. Chem.* **2009**, 52(14), 4454-65.

[111] Ekonomiuk, D.; Su, X. C.; Ozawa, K.; Bodenreider, C.; Lim, S. P.; Yin, Z.; Keller, T. H.; Beer, D.; Patel, V.; Otting, G. Discovery of a non-peptidic inhibitor of West Nile virus NS3 protease by high-throughput docking. *PLoS Negl. Trop. Dis.* **2009**, 3, e356.

[112] Krüger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *Chem. Med. Chem.* **2010**, 5(1), 148-58.

[113] Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated docking screens: A feasibility study. *J. Med. Chem.* **2009**, 52 (18), 5712–5720.

[114] Hartshorn, M. J.; Verdonk,M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, highquality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, 50, 726–741

[115] Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program:DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, 20, 601–619.

[116] Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. Molecular docking screens using comparative models of proteins. *J. Chem. Inf. Model.* **2009**, 49 (11), 2512–2527.

[117] Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The thermodynamics of protein-ligand interaction and solvation: Insights for ligand design. *J. Mol. Biol.* **2008**, 384, 1002–1017.

[118] Gilli, P.; Ferretti, V.; Gilli, G.; Borea, P. A. Enthalpy-entropy compensation in drug-receptor binding. *J. Phys. Chem.* **1994**, 98, 1515–1518.

[119] Das, S.; Krein, M. P..; Breneman, C. M. Binding affinity prediction with property-encoded

shape distribution signatures. *J. Chem. Inf. Model.* **2010**, 50 (2), 298–308.

[120] 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.* **2002**, 41, 2644–2676.

[121] Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, 46, 2287–2303.

[122] 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.

[123] Weill, A.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, 50, 123-135.

[124] Feldman, H. J.; Labute, P., Pocket similarity: Are α carbons enough? *J. Chem. Inf. Model.* **2010**, 50 (8), 1466–1475.

[125] De Franchi, E.; Schalon, C; Messa, M.; Onofri, F.; Benfenati, F. Binding of protein kinase inhibitors to synapsin i inferred from pair-wise binding site similarity measurements. *PLoS ONE* **2010**, 5(8), e12214.

[126] Reisen, F.; Weisel, M.; Kriegl, J. M.; Schneider, G. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *J. Proteome Res.* **2010**, ASAP.

[127] Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, 462, 175-181.

[128] Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotech.* **2006**, 24, 805–815.

[129] Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat

multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, 5(7), e1000423.

[130] Das, S.; Kokardekar, A.; Breneman, C. M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, 49 (12), 2863–2872.

[131] Sukumar, N.; Krein, M.; Breneman, C. M. Bioinformatics and cheminformatics: Where do the twain meet? *Curr. Opinion Drug Disc. Devel.* **2008**, 11 (3), 311-319.

[132] Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, 5 (4), 262-275.

[133] Mestres, J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Disc. Dev.* **2004**, 7 (3), 304-313.

[134] Oloff, S.; Zhang, S.; Sukumar, N.; Breneman, C.; Tropsha, A. Chemometric analysis of ligand receptor complementarity: Identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J. Chem. Inf. Model.* **2006**, 46 (2), 844-851.

[135] Xie, L.; Bourne, P.E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. USA* **2008**, 105, 5441–5446.

[136] Ren, J.; Xie, L.; Li, W. W.; Bourne, P. E. SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison, *Nucleic Acids Res.* **2010**, 38 (Web Server issue): W441-W444.

[137] Das, S.; Krein M. P.; Breneman, C. M. PESDserv: a server for high-throughput comparison of protein binding site surfaces. *Bioinformatics* **2010**, 26 (15), 1913-1914: http://reccr.chem.rpi.edu/Software/pesdserv/

[138] Martin, R. E.; Green, L. G.; Guba, W.; Kratochwil, N.; Christ, A. Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach *J. Med. Chem.* **2007**, 50, 6291-6294.

[139] Gloriam D. E.; Foord, S. M.; Blaney, F. E.; Garland, S. L. Definition of the G-protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J. Med. Chem.* **2009**, 52 (14), 4429–4442

[140] Sarah L. Kinnings and Richard M. Jackson, Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family, J. Chem. Inf. Model., 2009, 49 (2), pp 318–329

[141] Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. QSAR models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *J. Chem. Inf. Model.* **2009**, 49, 1974-1985.

[142] Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chem. Biol.* **2008**, 4, 682-690.

[143] Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, 50 (8), 1418–1431.

[144] Belongie, S.; Malik, J.; Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, 24, 509–522.

[145] Roth, B. L.; Sheffler, D. J.; Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Disc..* **2004**, 3, 353–359.

Figure Captions

Figure 1: Wiener number W is the total distance between all carbon atoms in a molecule. The smaller this number, the more compact the molecule. W is obtained by multiplying the number of carbon atoms on one side of any bond by those on the other side, and summing the result for all bonds. W can also be obtained by adding all the elements of the graph distance matrix above the main diagonal. Connectivity index χ is constructed from the row sums R_i and R_j of the adjacency matrix using the algorithm $\chi = \sum_{ij}(R_iR_j)^{-1/2}$ with each bond ij having a contribution $(R_iR_j)^{-1/2}$. χ is a bond additive quantity where terminal CC bonds are given greater weight than inner CC bonds.

Figure 2: PEST Property-encoded ray tracing. A property-encoded surface is subjected to internal ray reflection analysis. A ray is initialized with a random location and direction within the molecular surface and reflected throughout inside the electron density isosurface until the molecular surface is adequately sampled. Parts of the molecular van der Waals surface are shown encoded by a property, such as the electrostatic potential (EP), and other other parts are shown cut away to reveal the ray traces. Molecular shape information is obtained by recording the ray-path information, including segment lengths, reflection angles and property values at each point of incidence.

Figure 3: PEST shape-property distribution for electrostatic potential (EP) and ray length, and corresponding 2-D histogram signature.

Figure 4: Traditional regression approaches that minimize the training errors $\Sigma_i(y_i - f(x_i))$ lead to over-fitting in HTS data with noise; *i.e.*, to a “perfect” prediction for the training set, but poor predictions for unknown data.

Figure 5: Artificial Neural Networks for biological activity prediction. A non-linear transfer function (generally of sigmoid form) is applied at each node to determine the output value of that node from the input signals received from other nodes connected to it. The weight of each connection (“synapse”) is optimized during the learning phase to produce the best response for the training data.

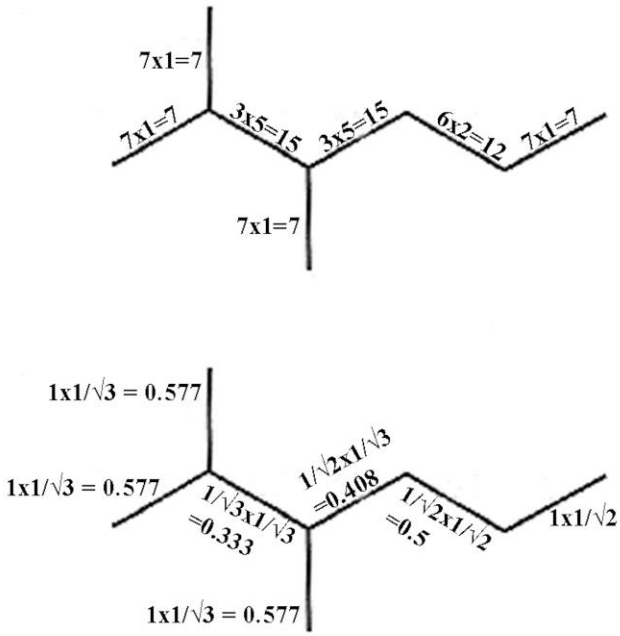
Figure 6: Protocol for predictive QSAR modeling and validation.

Figure 7: Support Vector Machine classification. The goal of the machine is to find the best classification of the data (filled and open circles) my maximizing the separation or margin between the “support vectors” (dotted lines) on either side of the classification hyperplane.

Figure 8: Support Vector regression minimizes the generalization error, defined as the sum of the training error and the model complexity. Minimizing ||w|| controls the capacity of the function, while the parameter C controls the tradeoff between error and capacity. The linear penalty or loss function L_ε is applied to everywhere outside the margin or ε-tube. SVM models avoid over-fitting by controlling the model complexity.

Figure 9: Property Encoded Shape Distributions (PESD): Conversion of property distribution on surfaces to a string of numbers or signatures. A large number of randomly selected pairs of points from the surface are binned by distance and property combinations to construct PESD signatures. The similarity between two binding sites is calculated from the similarity between the corresponding PESD signatures

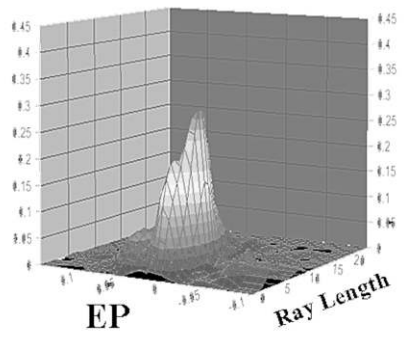
Figure 10:. Binding site representations of of 1b55 and 1btn show low sequence conservation. However, the EP mapped surfaces and corresponding PESD signatures are very similar [130].



Wiener Number
W=7+7+15+7+15+12+7
= 70

Connectivity Index
$\chi = \sum 1/\sqrt{(R_i R_j)}$
= 3.681

Figure 1



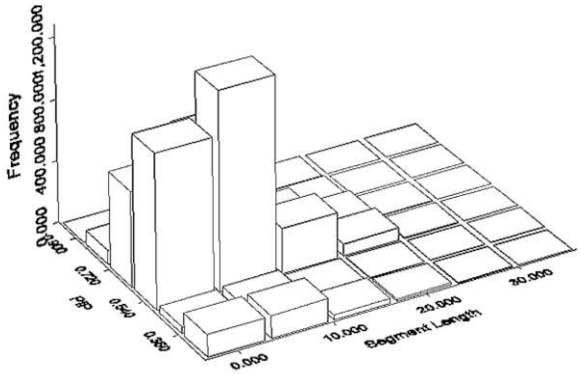


Figure 3

Figure 4

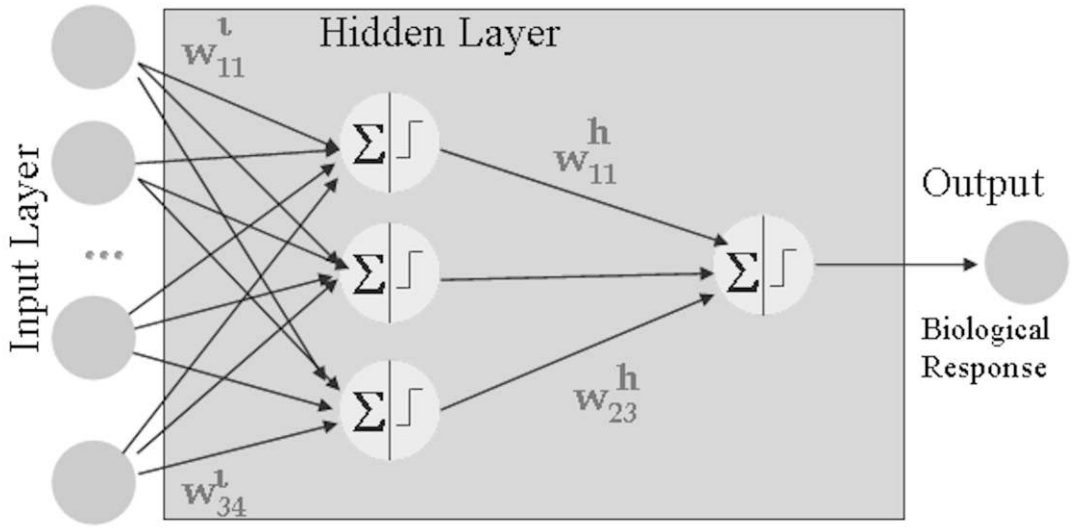


Figure 5

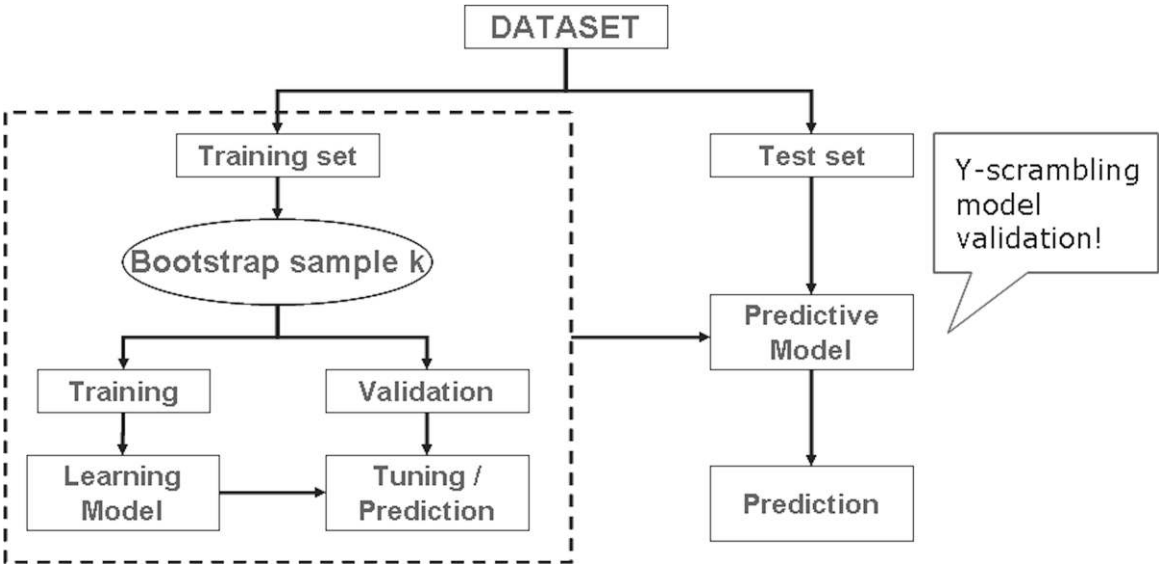


Figure 6

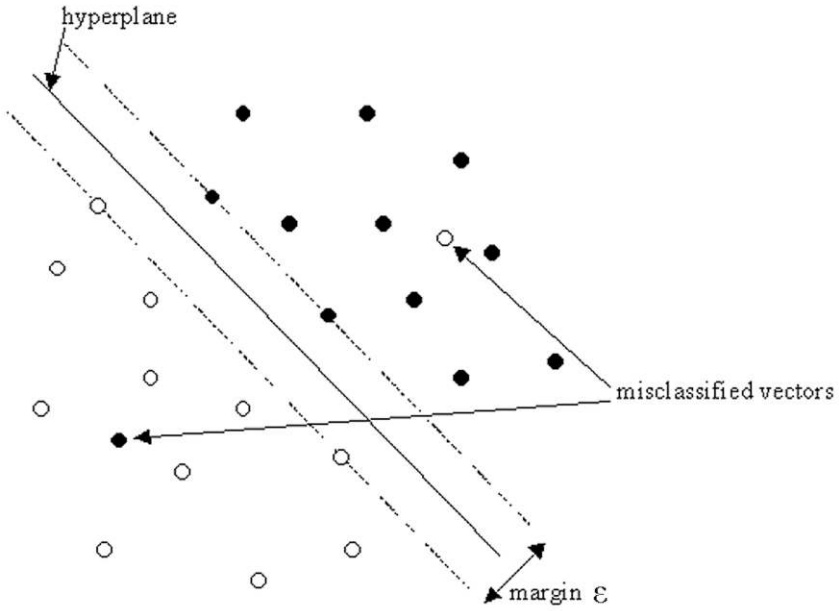


Figure 7

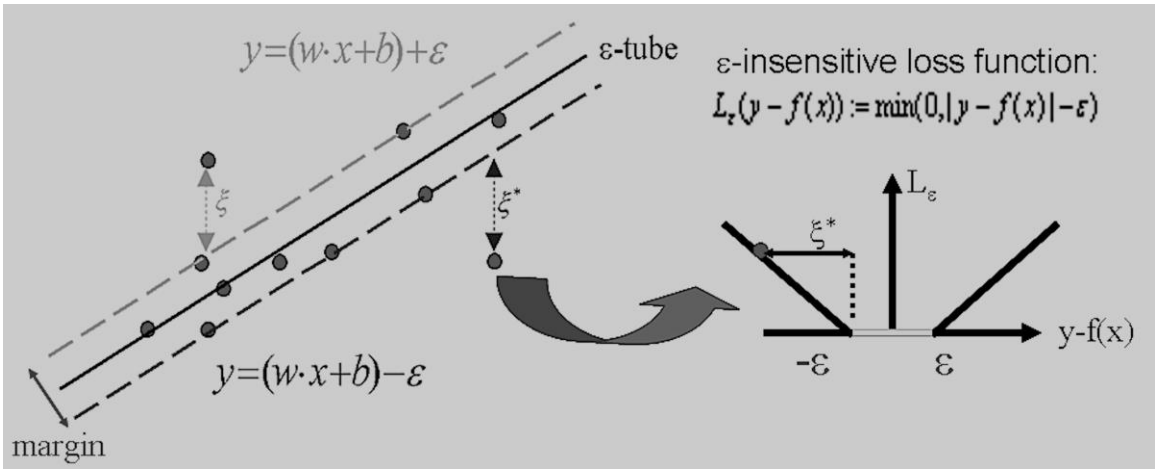


Figure 8