Variable selection and model validation of 2D and 3D molecular descriptors

Anthony Nicholls[a,]*, Norah E. MacCuish[b] & John D. MacCuish[b]
[a]*OpenEye Scientific Software, Inc., 3600 Cerrillos Rd. Suite 1107, Santa Fe, NM 87507, USA;* [b]*Mesa Analytics & Computing, LLC, 212 Corona St., Santa Fe, NM 87501, USA*

Received 4 May 2004; accepted in revised form 26 August 2004

Key words: cluster analysis, electrostatics, field overlaps, hypothesis generation, *k*-fold cross-validation, molecular shape, QSAR, structural fingerprints, Tversky

Summary

We have found that molecular shape and electrostatics, in conjunction with 2D structural fingerprints, are important variables in discriminating classes of active and inactive compounds. The subject of this paper is how to explore the selection of these variables and identify their relative importance in quantitative structure–activity relationships (QSAR) analysis. We show the use of these variables in a form of similarity searching with respect to a crystal structure of a known bound ligand. This analysis is then validated through *k*-fold cross-validation of enrichments via several common classifiers. Additionally, we show an effective methodology using the variables in hypothesis generation; namely, when the crystal structure of a bound ligand is not known.

Introduction

Quantitative structure–activity relationships (QSAR) analysis assumes a relationship between the chemical properties of a molecule and its physical behavior, such as binding to a target protein. As a result, researchers have tried to encode physicochemical properties and structure in so-called molecular descriptors. Finding a combination of such descriptors or variables that best reveals activity relationships across a wide range of ligand/target complexes is an area of ongoing research. This problem is confounded by the fact that ligand/target complexes can have numerous binding modes. It is well known that 2D chemical structure, molecular shape, and electrostatics all play a significant role in the binding of ligands to targets for most complexes, and should therefore correlate with the activity response found in screening compounds. Hence shape, electrostatic, and 2D structural descriptors that are robust and efficient should in principle provide the necessary ease of use and discriminatory power to fully reveal activity relationships.

Two-dimensional structural fingerprint methods [1–3] and chemical properties [4,5] are manifold, but researchers have also created many forms of 3D descriptors containing shape and electrostatic information and many methods of using such variables [6–8]. All have been used in QSAR analysis. Difficulties with these methods have been various. For instance, CoMFA does not generalize well beyond single chemical series and the alignment of active molecules can affect the accuracy. Pharmacophore and pharmacophore-like methods have no formal theory of comparison; i.e., there is no good way of understanding the statistical significance of a given model. In addition, pharmacophoric elements require ‘weightings’ that have no physical basis and hence are not easily transferable. Other methods (BCUTS, surface SOMs) have much the same reduction of 3D information found in 2D molecular fingerprints, and are thus incomplete in their molecular

*To whom correspondence should be addressed. Tel.: +1-505-473-7385; Fax: +1-505-473-0833l; E-mail: anthony@www.eyesopen.com description. In general, 3D methods are slow and limited in the scale of problems they can address.

3D methods also have particular problems, most notably the multiple conformer issue. Aligning shapes efficiently for comparison or electrostatic analysis is problematic. Accurate charges and good electrostatic models are needed for electrostatic comparisons. Additional problems can arise due to incorrect tautomers, ionization states, or mistakes in atom typing due to incomplete or incorrect chemistry perception. The latter can be particularly acute in the perception of aromaticity. The authors of the prediction method xLog P [9] considered furan non-aromatic and had a significant discrepancy with the Log P from experiment. This discrepancy disappears if furan is more appropriately considered aromatic. However, formal descriptions of the shape of molecular conformers that have been refined with modern force-fields, set with high-quality charges, utilizing a method of accurately calculating the potentials in solution, should have a decided advantage over previous approaches because these are the variables that are intimate to molecular interaction. Using these descriptors and their respective comparative measures with a drug-like, feature-based, 2D fingerprint such as the Molecular Design Limited 320 keys based fingerprint [10] presents a formidable combination of three relatively orthogonal variables for discrimination, or regress, on activity.

A crucial aspect that separates this work from others is that it is based on formal theory of molecular shape comparison. This theory is general, extensible and parameter free. It needs no tuning for each new application, as it requires no training. It shows that shape, as defined in this paper, is a fundamental molecular property and that shape difference forms a metric space. Lack of a theory of this nature need not negate the usefulness of a 3D method – heuristics and hypothesis generation are common enough in chemistry – but it limits the transferability, the reliability, and the generalization of the method, for instance, to other physical phenomena. In particular, when a heuristic fails there is no mechanism of assigning value to this failure, i.e., is this just a singleton not caught by the *ad hoc* collection of rules, or does the failure say something about the model and how it is being applied? What separates a theory from a heuristic is the value of negative information. Failures need to mean something with respect to the theory. We show several examples of 'good' failures that provided useful biophysical information in the Results section.

There are several important issues to settle in developing useful models with shape and electrostatics. The first is to reliably construct the conformer search space with a robust measure of internal energy: high-energy conformers are unlikely to contribute meaningful information and can swamp a signal from low-energy shapes. The second is to quantify and ameliorate the sensitivity of the electrostatic similarity to shape alignment. The third is to design an optimal combination of shape and electrostatic comparisons. With effective methods for these issues in hand, shape and electrostatic descriptors can address two forms of traditional QSAR analysis. The first is similarity searching to a bound crystal ligand so as to produce a predictive model; the second is hypothesis generation without structural data, e.g. assay data alone.

In the Theoretical basis section, the shape, electrostatic, and 2D MDLI descriptors, and their respective measures, will be discussed in detail. An overview of the forms of unsupervised and supervised learning techniques, used in the similarity searching and hypothesis generation, will be addressed. Methods for conformer search space, conformation and electrostatic matching, and the optimal association between conformers and electrostatic comparisons will be presented in the Experimental methods section, as are overall protocols for the use of shape and electrostatic descriptors and measures used in similarity searching and hypothesis generation. Finally, similarity searching results for Cox2 and Progesterone, and the hypothesis generation for the assay data of Dopamine and Calcium ion channel are presented in the Results section.

Theoretical basis

Shape

The theoretical basis for molecular shape analysis in this work derives from the concept of volume overlap, as first applied by Masek et al. [11]. Molecules are typically represented visually and hence conceptually as a set of N overlapping hard spheres.

If $\chi_i(\boldsymbol{r})$ is a function equal to one inside sphere i's radius and zero outside, then the function:

$$\chi(\boldsymbol{r}) = 1 - \prod_{i=1}^{N}(1 - \chi_i(\boldsymbol{r})) \tag{1}$$

describes a 'volume' function for a set of N spheres. The vector notation here is that $\boldsymbol{r}$ is a (x,y,z) position in space. If the spheres do not overlap, Equation 1 reduces to:

$$\chi(\boldsymbol{r}) = \sum_{i=1}^{N} \chi_i(\boldsymbol{r}).$$

This expression is now simply the sum of the volumes of the individual spheres. Otherwise its expansion includes terms that correct for over-counting the overlaps of spheres. These overlaps can be of order $N-1$, and hence very complex. However, for most hard-sphere representations of molecules, only those of first-order contribute significantly to the volume representation, because bond lengths are commensurate with atomic radii. As such, it is relatively straightforward to calculate the volume of a hard-sphere molecule from this formula.

Given a functional form of the volume of a molecule, there follows the essential measure of shape comparison: volume overlap.

$$O_{\mathrm{A,B}}(\boldsymbol{q}^{\mathrm{A}}, \boldsymbol{q}^{\mathrm{B}}) = \int\int\int \chi^{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{q}^{\mathrm{A}})\chi^{\mathrm{B}}(\boldsymbol{r}, \boldsymbol{q}^{\mathrm{B}})d\boldsymbol{r}. \tag{2}$$

The integral here is over all space but, for solid-sphere functions, this reduces to being over a volume that includes molecules A and B. The variables $\boldsymbol{q}^{\mathrm{A}}$ and $\boldsymbol{q}^{\mathrm{B}}$ represent orientation variables, rotations and translations to each molecule, which do not affect either's shape, merely the relative overlap. Fixing either $\boldsymbol{q}^{\mathrm{A}}$ or $\boldsymbol{q}^{\mathrm{B}}$ and optimizing Equation 2 with respect to the other leads to the concept of the maximal overlap:

$$O_{\mathrm{A,B}} = \max\{O_{\mathrm{A,B}}(\boldsymbol{q}^{\mathrm{A}}, \boldsymbol{q}^{\mathrm{B}})\}.$$

If we know how to calculate a molecular volume function $\chi(\boldsymbol{r})$, we can apply this knowledge to the construction of an overlap function, the optimization of which gives a volume overlap $O_{\mathrm{A,B}}$ that is a measure of the *global* similarity of the shapes of two molecules. In fact, one can calculate a shape 'distance' or proximity, thus:

$$D_{\mathrm{A,B}} = \sqrt{O_{\mathrm{A,A}} + O_{\mathrm{B,B}} - 2 * O_{\mathrm{A,B}}}.$$

This quantity is a true metric quantity. It is always positive, only zero when two volumes are identical and obeys the triangle inequality:

$$D_{\mathrm{A,B}} + D_{\mathrm{B,C}} \geq D_{\mathrm{A,C}} \geq |D_{\mathrm{A,B}} - D_{\mathrm{B,C}}|.$$

The proof of this conjecture is straightforward but, to our knowledge, has not been previously published. The proof comes in two parts. The first is that given any arrangement of three bodies the triangle inequality must hold for the field distance between each. This must be true because the field distance is the difference in two integrals, each of which can be approximated, to arbitrary accuracy, by an ordered list of numbers, and thus a vector representing a discretization of the integral. Since the triangle inequality holds for two vectors of arbitrary dimension, the first part is then proven. The second part derives from the following *gedanken*: Optimally align A with B, and C with B. Given the first part of the proof:

$$D^{\mathrm{opt}}_{\mathrm{A,B}} + D^{\mathrm{opt}}_{\mathrm{B,C}} \geq D_{\mathrm{A,C}}.$$

But

$$D_{\mathrm{A,C}} \geq D^{\mathrm{opt}}_{\mathrm{A,C}}$$

by definition. Therefore:

$$D^{\mathrm{opt}}_{\mathrm{A,B}} + D^{\mathrm{opt}}_{\mathrm{B,C}} \geq D^{\mathrm{opt}}_{\mathrm{A,C}}.$$

The lower bound of the triangle inequality follows from the upper bound and hence the conjecture is proven. The implications for shape comparison are considerable and will be noted below. In particular, it implies that the shape of a molecule can be thought of as a point in some high-dimensional space, and therefore to possess an absolute property, although we always think of it as a relative concept. Other work has pursued this conjecture [Grant et al., in preparation] and suggests that the dimensionality of molecular shape of small molecules (<32 heavy atoms) is around 20. For our work what is important is that shape distance is both a fundamental property and a measurable one. As will be seen, shape clustering is very powerful, presumably reflecting its nature as a true distance.

If distance in 'Shape Space' is not a familiar concept, the concept of a Tanimoto similarity measure is very common in chemical structural space. The following formula yields a 'Shape Tanimoto' with a value between one and zero, one being identical, zero completely dissimilar.

$$T_{A,B} = O_{A,B}/(O_{A,A} + O_{B,B} - O_{A,B}).$$

The 'Shape Tanimoto' can never be equal to zero: molecules can always be made to overlap to some extent. It can be equal to one if A is identical to B. An asymmetric Shape Tversky measure that can be used to evaluate how well one molecule fits within another can also be defined:

$$TV_{A,B} = O_{A,B}/(\alpha O_{A,A} + \beta O_{B,B} - O_{A,B}).$$

The Tanimoto measure can have metric properties (or rather its difference from unity, namely the Soergel measure, may have such [12]). However, the Tversky measure is not a metric under any simple transformation, if α does not equal β, and α and β are both non-zero. The advantage of Tversky measures is the representation of partial shape similarity.

Electrostatics

Effective molecular electrostatics requires two components: a good charge model and a good method of using those charges. In this work there is one charge model, that arising from the molecular force-field MMFF94 [3, 4]. MMFF94 charges are derived from parameterization of functional groups common in medicinal chemistry. We have found MMFF94 more reliable than other bond-increment schemes [15].

The method of using the atom-centered charges calculates a grid of potentials via the solution of the Poisson equation. This, then, includes the influence of solvent on the potential distribution and has the effect of dampening the long-range effects of individual charges. This grid of potentials is then used to calculate field overlaps similar to that for the steric field function $\chi(\boldsymbol{r})$:

$$EO_{A,B} = \iiint \phi^{A}(\boldsymbol{r})\phi^{B}(\boldsymbol{r})\mu^{A}(\boldsymbol{r})\mu^{B}(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \qquad (3)$$

$\mu(\boldsymbol{r})$ is a masking function that dampens the electrostatic field values, $\phi(\boldsymbol{r})$ in the interior of the steric volume. We damp internal potentials because they are not relevant for the comparison of molecular interactions. Typically $\mu(\boldsymbol{r})$ has the form:

$$\mu(\vec{r}) = 0; \quad \chi(\vec{r}) > 0.5,$$
$$\mu(\vec{r}) = 1; \quad \chi(\vec{r}) \leq 0.5.$$

Notice that $EO_{A,B}$ has no dependence on any rotational or translational parameters because those have already been determined by the steric alignment. A concern about EO is that it is not necessarily finite, i.e., Equation 3 can fail to converge. The long-range electrostatic field of a charged molecule is that of a monopole centered at the charge-weighted average atom center. Two charged molecules therefore have a product field that falls off as $1/\boldsymbol{r}^2$, i.e. long-range contributions of $O(1)$. This can be avoided if the masking function $\mu(\boldsymbol{r})$ is allowed to mask off long-range as well as short-range potentials. As a practical matter, the 'shape' of electrostatics close to the molecule is all that concerns us as a measure of similarity, as, for instance, in CoMFA or ESP fitting of atomic charges to *ab initio* field potentials.

The Electrostatic Field Tanimoto, $\mathrm{ET}_{A,B,}$ can be calculated from the formula:

$$\mathrm{ET}_{A,B} = EO_{A,B}/(EO_{A,A} + EO_{B,B} - EO_{A,B}).$$

An interesting oddity of this measure is that it can be negative. Suppose the field of molecule B is exactly opposite to that of A, thus positive where it is negative and negative where it is positive. Then we have:

$$\mathrm{ET}_{A,B} = \mathrm{ET}_{A,-A} = -EO_{A,A}/(EO_{A,A} + EO_{A,A} - EO_{A,A}) = -1/3,$$

where the minimal value of the ET is minus one third. If the fields of molecules A and B are the same, then the measure equals one.

2D structural fingerprints: MDL 320 key fingerprints

Characterizing chemical structures in binary form based on 2D representations facilitates many tasks of chemical informatics. Such binary representations are called *molecular fingerprints.* Molecular fingerprints were first developed by chemical information systems (CIS) companies for efficiency enhancements in chemical database queries, and not necessarily for QSAR studies.

Several methodologies exist for chemical binary representations. For example, Daylight CIS Inc. [1] fingerprint is often referred to as a *path* or subgraph approach. This amounts to a unique sub-graph matching of the graph representation of the chemical structure. In the Daylight algorithm the fingerprint is 'learned' from the structures themselves. A molecular fingerprint is generated from a hash of all the unique connection paths (sub-graphs) up to a maximum size, typically 8, into a fixed length bit string. Fingerprints may be folded to decrease the length and increase the bit density. Typical sizes for Daylight fingerprints are 512 or 1024 bits in length, but any power of two can be generated.

Molecular Design Limited Information Systems create a key-based fingerprint, where the 'key' is a chemical sub-structure. This fingerprint uses a pre-defined set of definitions and creates fingerprints based on pattern matching of the structure to the defined 'key' set. This approach relies on these definitions to encapsulate the molecular descriptions *a priori* and does not 'learn' the keys from the chemical dataset. The MDL original public key set is 166 keys, and their private key set is comprised of 966 keys. Their recent publication of 'drug-like' keys contains a subset of 320 keys from their 966-key set and these are designed with similarity searching and QSAR in mind [10].

The Mesa Analytics & Computing, LLC's implementation of the MDL 320 key fingerprints contains a SMARTS pattern or a set of SMARTS patterns for each key [16]. The fingerprints generation consists of SMARTS pattern matching against the chemical dataset using the algorithm contained in OEChem [17].

Statistical learning theory: similarity searching

Comparing a bound crystal ligand with other active and inactive ligands is a form of similarity searching [18]. Ordinarily, similarity searching is performed with one variable, such as binary fingerprints, or a set of variables that are transformed to produce a ranking (e.g., BCUTS). Typically the searching is in an experimental setting where there are a great many more inactives than actives. The class separation is often so poor that these one-dimensional searches are reported in terms of enrichments, where enrichment is defined as the percent increase in actives over the percent found in the data sample as a whole. The enrichments are reported in terms of what multiple of the actives is found given the top x percent of the ranking vs. the original prior percentage of the data set (or what would be found by pure random selection of the data). Ranking percentages of 0.1, 1, 5, and 10 are commonly reported with an enrichment curve showing all percentages compared to the random selection curve as a base-line measure.

There are a number of problems with this approach. Rarely are such results reported in terms of cross-validation or is a cost function defined, such as possible weightings given to the true and false positives. Rank percentage cutoffs define implicit cost functions, but no explicit coefficients for the cost function in terms of true and false positive rates, for example, are given. Without cross-validation of the reported enrichment values it is difficult to assess the variables or model used to generate them. However, k-fold cross-validation of enrichments suffers from wide variance for two reasons: the test samples' ratio of actives to inactives differs from the prior ratio of the training set to which they are compared; and, the enrichments values are a ratio of a ratio whose distributions typically have the long wide tails of a Cauchy distribution. Additionally, enrichment can be infinite, if the test set classes are perfectly separated by the model.

Thus, if the class separation is at least modest and there is more than one variable, rather than enrichments, more appropriate measures of cross-validated classification model performance with uneven class sizes can be used. As an example, k-fold cross-validated accuracy, normalized to uneven class sizes such as geo-mean2 [19], more appropriately captures the mean and variance of the model performance. Cost function coefficients can easily be added to these models, but without a company or industry-wide standard as to the acceptable error rates of false negatives and false positives, other than some general notion that false negatives are often discounted, we show the results in terms of equal percentage weighting of false negatives and false positives.

Statistical learning theory: clustering and hypothesis generation

Unsupervised learning algorithms, such as clustering algorithms, attempt to find groupings in data, and are often used in exploratory data analysis or segmentation studies that determine *ad hoc* classes. There are numerous heuristics to determine the 'best' number of groups, often varying from algorithm to algorithm. Regardless of the application, it is typically considered best not to rely on a single such algorithm or heuristic to determine the best number of groups (k), but to explore the results of a suite of clustering algorithms [20] armed with any expert knowledge at hand, in order to make a determination if the groups found are reasonable both quantitatively and qualitatively.

Experimental methods

Shape overlay: ROCS

The theoretical description of shape overlay given above requires an effective method of finding $O_{\mathrm{A,B}}$, the maximal overlap between two volumes A and B. Masek [11], in his original work, used the gradients from spherical intersections to search for the optimum overlap. The procedure required many attempts from different starting orientations to find a global minimum, because hard-sphere functions are discontinuous, leading to a convoluted, 'brittle' energy surface. Shortly afterwards Grant and Pickup suggested this difficulty could be overcome by replacing the χ function for a hard sphere with a Gaussian.

$$\chi(\boldsymbol{r}) = pe^{-\gamma(r-c)^2}.$$

Here c is the center of the atom and γ and p are the Gaussian width and height that determine the volume, where the volume equals the integration of this function over all space:

$$V = p^* \sqrt{(\pi/\gamma)^3}.$$

Since V should also equal $4\pi R^3/3$, the volume of a sphere of radius R, γ and p are not independent:

$$\gamma = \pi(3p/4\pi R^3)^{-2/3}.$$

In a seminal paper [21], Grant and Pickup showed that, by a judicious choice of Gaussian height, the volume of molecules could be calculated to within 0.1%. This was a remarkable result for several reasons. The Gaussian χ function is quite different from the hard-sphere function it replaces: it is smooth; it extends to infinity and is not of unit height, confounding the interpretation of χ with a value of either 0 or 1. Secondly, p is universal for all atoms, only γ varies from atom to atom, depending on the hard-sphere radius. Thirdly, the product formula of Equation 1 for the exact total volume is now quite tractable. Instead of this formula producing increasingly complex analytic forms for the intersection of a set of spheres, all such intersections are now merely products of Gaussians. As products of Gaussians are also Gaussians (the proof is simply that the sum of quadratic functions of a variable is also quadratic), intersections of any order are trivial to construct and evaluate. Grant found that using Gaussians not only was the energy landscape much smoother, allowing much more rapid and reliable convergence, but that the final overlays resembled closely those found from the hard-sphere method. In addition, he found that the global minimum could usually be found from one of four starting points: the four degenerate overlays of the inertial axes of the molecule.

The program ROCS (Rapid Overlay of Chemical Structures) [22] is an implementation of the Grant and Pickup work. It represents the molecular volume by a set of Gaussians, and attempts, by numerical optimization, to maximize the overlap of two molecules from a set of rational starting positions. Optimization is performed in quaternion space and can be performed using either analytic forms or grid-based approximations. It typically ignores hydrogens, as their volumes are not significant to shape (they do not alter the orientation obtained). It can evaluate $O_{\mathrm{A,B}}$ for a pair of typical drug-sized molecules in about a millisecond on a modern Intel/AMD processor. As such, ROCS has become a useful tool to search conformationally expanded corporate collections for molecules similarly shaped to a known active. ROCS has introduced several short-cuts to the Grant/Pickup prescription to increase throughput, for instance only keeping the zero order Gaussians; i.e., those representing atoms, not overlaps of atoms, and setting all heavy atoms a single radius. It also extended the Grant prescription in some areas, for instance molecules of high symmetry require more starting points to reliably find the global minimum. Finally, ROCS reports of the Shape Tanimoto and Shape Tversky, a feature that has been predominantly exercised in this work. An example of a ROCS overlay can be found in Figure 1.

Electrostatic Tanimoto: EON

The complementary program to ROCS for electrostatics is EON. Given a pair of molecular structures and an alignment, for instance from ROCS, EON calculates the electrostatic field around each using the ZAP toolkit. ZAP is a Poisson–Boltzmann (PB) solver, such that it solves the PB equation wherein molecular interiors are assigned one dielectric, and exteriors (typically water) another, thereby taking into account solvent polarization. Charges are either assigned by the program (typically MMFF charges) or read from the molecule files. Fields are typically extended out from the molecular surface by 6 Å or more and are enumerated on a regular lattice, or grid, with a spacing of about 1 Å. The fields are then masked by a $\mu(\boldsymbol{r})$ (Equation 3), derived from a ROCS-type shape function $\chi(\boldsymbol{r})$ produced by the sum of atom-centered Gaussians, and then multiplied together to form the product $EO_{A,B}$. Typical calculation times per pair of molecules are 0.1–0.2 s.

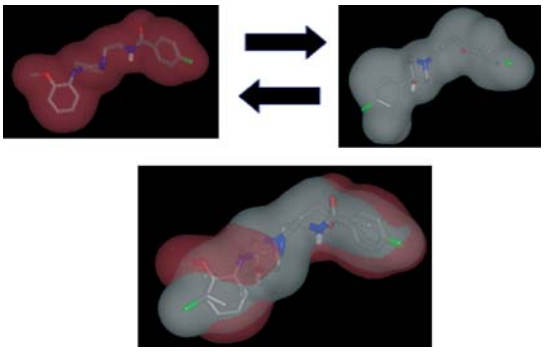

Figure 1. The Shape Tanimoto value between these two structures is 0.781. The shape overlap is portrayed by the exposed color of one structure vs. that of the other.

Structure generation: OMEGA

OMEGA is a rotor-driven conformer generation program. A molecular structure is parsed for rotatable bonds, fragmented by such bonds and each fragment reassembled into a complete molecule based upon defined torsion angles. If no initial molecular structure is provided, e.g., if only a SMILES string or a 2D-connection table is available, an initial structure is calculated using a distance bounds method [23]. The energy of each structure is evaluated with a force-field (here the Merck force-field, MMFF94) and low-energy structures are retained. In addition, root-mean-square (RMS) can be used to remove structures that are similarly shaped so that the collection represents a diverse sampling of low-energy conformations. Of particular importance is that OMEGA is a deterministic and exhaustive

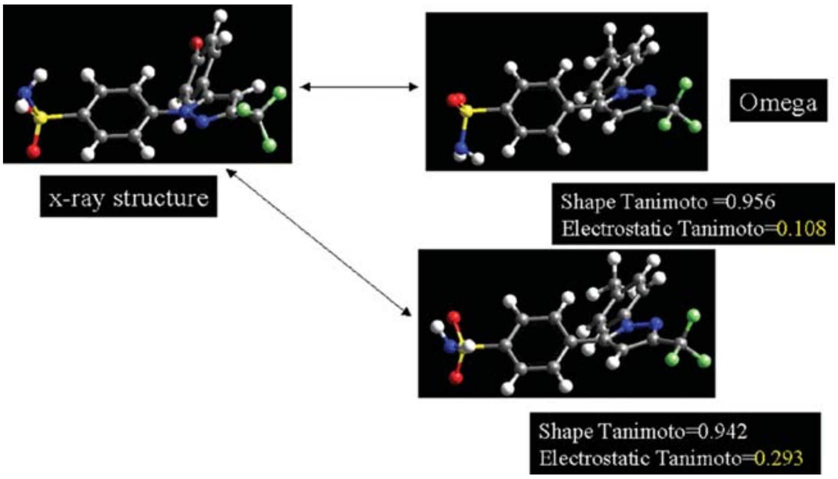

Figure 2. An X-ray structure compared to two conformations generated with OMEGA. Negligible Shape Tanimoto score changes by rotor rotations can result in significant Electrostatic Tanimoto score changes.

algorithm. There are no stochastic components that increase the likelihood of an important conformation not being present in the final ensemble. It has been shown that OMEGA reliably reproduces crystal structures of ligands [24].

Dataset selection

All X-ray crystal structures were obtained from the Protein Data Bank [25]. Active ligands and decoys were selected from the Wombat Database (WD) [26]. Highly active ligands have reported activities greater than 10 nM, moderately active ligands have activities greater than 1 μM and less than 10 nM, and weakly active ligands have non-zero activities less than 1 μM. Charged structures were removed, for reasons discussed above. The importance of the WD cannot be over-emphasized. Our early work in this area was hampered by the occurrence of true actives in the decoy set. This is a natural consequence of choosing decoys from 'drug-like' sets, e.g. MMDR. The WD allows us to choose compounds that are active 'drug-like' molecules towards specific ligands but are not reported to have activity to a particular target of interest, because of the database's careful annotation of activity. This was a significant contributing factor to the strength of the results.

Parameter selection

Conformational structures were generated from SMILES input with OMEGA version 1.5b. Shape Tanimoto values were calculated using ROCS version 2.0.0 and Electrostatic Tanimoto values were calculated using EON version 0.2b. The single conformer selected for each comparison was chosen as the one with the best geometric mean of the Shape Tanimoto and Electrostatic Tanimoto. The reason the geometric mean was implemented is due to the fact that small changes in Shape Tanimoto can result in large changes in Electrostatic Tanimoto, as described in Figure 2. For example, rotating a polar and terminal rotor group can produce a small Shape Tanimoto change but a large variance in Electrostatic Tanimoto.

In Figure 3 a comparison is shown between selection of conformers with the best geometric mean of the Shape and Electrostatic Tanimoto values or the conformer with the largest Shape Tanimoto. Figure 3 reveals that selection of a conformer with a slightly smaller Shape Tanimoto, can in fact lead to a significant change in the Electrostatic Tanimoto. The geometric mean method was implemented for all conformation selections for all datasets.

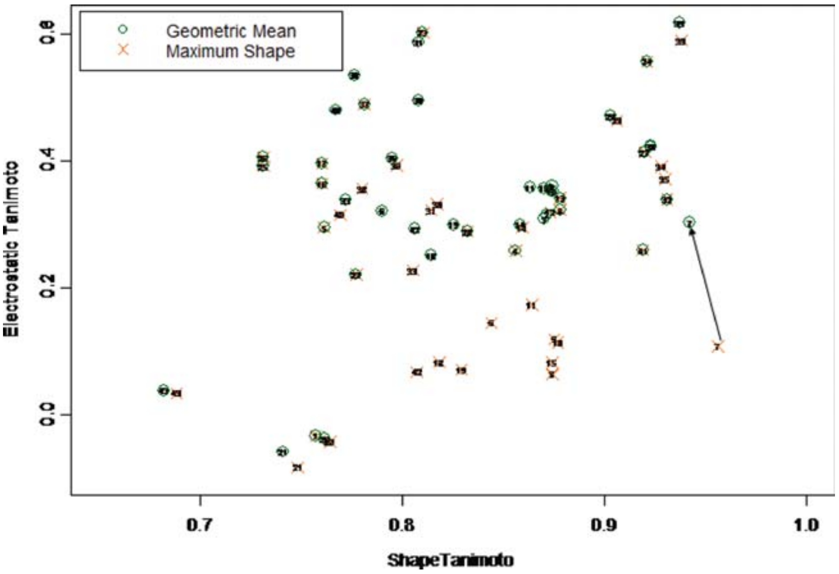

Figure 3. Comparison of two methods for conformer selection. Selecting the conformer with the best Shape Tanimoto vs. selecting the conformer with the best geometric mean of the Electrostatic Tanimoto and the Shape Tanimoto. Geometric mean = maximum shape 40% of the time. Sixty percent of the time, using the geometric mean, the Electrostatic Tanimoto is larger with only a small change in Shape Tanimoto.

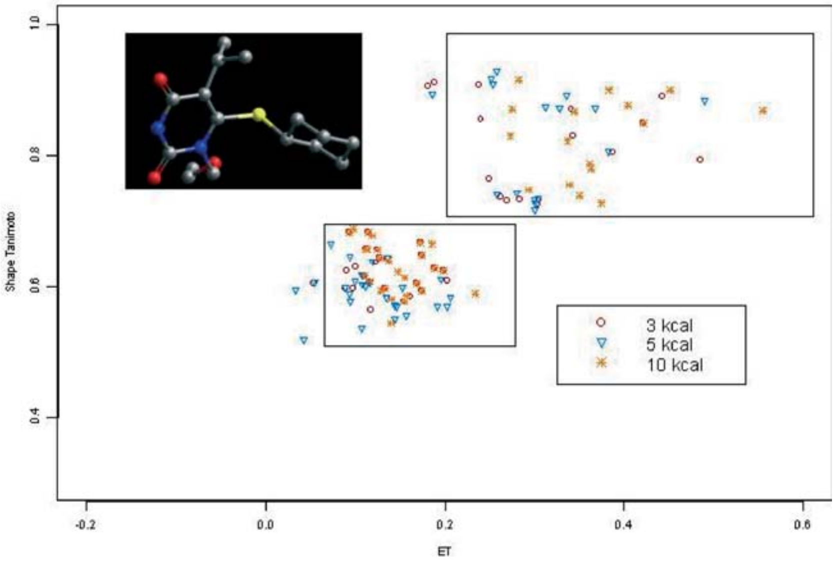

Figure 4. Shape Tanimoto vs. Electrostatic Tanimoto comparison for an X-ray structure vs. conformations generated at three different energy windows in OMEGA, 3, 5, and 10 kcal/mol. Best conformers were chosen using the geometric mean selection criteria.

A maximum of 100 conformations per structure were generated with OMEGA version 1.5b using an energy window of 5 kcal/mol and an RMS of 0.6 Å. This energy window size was chosen from the analysis illustrated in Figure 4. Conformations from SMILES reported to have activities with HIV-RT were generated with OMEGA at three different energy windows: 3, 5, and 10 kcal/mol. The Shape and Electrostatic Tanimoto values were calculated and the 'best' conformer was chosen using the geometric mean method described above. The comparison was made to an X-ray structure of a known HIV-RT active ligand. In Figure 4 the results are plotted. The distribution of the 5 kcal/mol results and the 10 kcal/mol results are not significantly different from the 3 kcal/mol results. This minimal energy window is, however, quite narrow and, in practice, we opted for a 5 kcal/mol margin.

In Figure 5, the RMS vs. the Electrostatic Tanimoto values are plotted for a set of OMEGA conformers generated from the SMILES of X-ray structures of known HIV-RT ligands bound to HIV-RT. The bound ligands 1c1c, 1rti, 1c1b, and 1rt2 are from the Protein Data Bank [25]. The SMILES were generated from the PDB files with Babel2[1]. MKC-442 is the target structure and the Electrostatic Tanimoto scores for a 'good' and a 'poor' conformer vs. MKC-442 are given in Figure 5. RMS values in the figure are between the conformer and the coordinates of their respective X-ray structure. One can see that for a 'good' Electrostatic Tanimoto score, one also has found a conformer that is close to the X-ray structure of the bound ligand. For a conformer with a 'poor' Electrostatic Tanimoto score, the conformer has a large RMS, and is thus far away from an active conformation. The critical threshold would seem to be of an RMS of 2.0 Å vs. an Electrostatic Tanimoto of 0.2. The former is what would typically indicate a failure in reproducing the bound conformation and hence, in what follows, we take the latter, an Electrostatic Tanimoto of 0.2, as significance threshold for activity, i.e. the minimum we would expect given a reasonable reproduction of binding mode.

Clustering methods

In this study, Ward's, Group Average, and Complete Link hierarchical algorithms [27]; and variants of the Taylor–Butina (TB) [28,29] leader algorithm are used to form a consensus as to the most likely groupings. Thresholding rather than level selection techniques were used to find appropriate cutoffs between the hierarchical algorithms and the TB. The groups found for Group Average, Complete Link, and TB can all be related directly to the similarity measures used. For

[1] Babel2: an example program of Oechem [17].

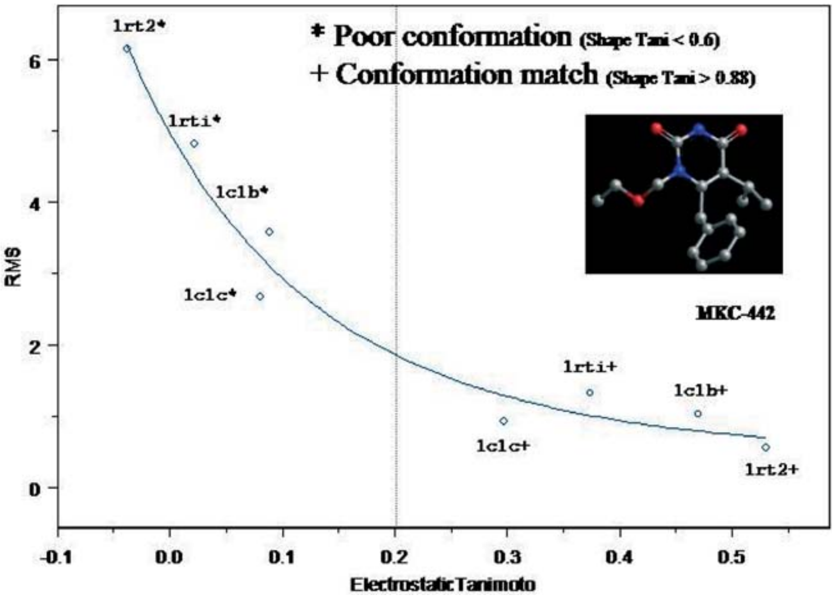

Figure 5. RMS vs. Electrostatic Tanimoto between OMEGA conformations generated from SMILES for four X-ray structures compared with MKC-442.

clustering using Shape Tversky, a consensus was formed using an asymmetric version of TB [30], and an asymmetric hierarchical clustering algorithm [30, 31]. All of the clustering algorithms used in the study were from Mesa Analytics & Computing's *Grouping Module*.

Statistical methods

For reasons of data availability and quality, classification was chosen over regression. The data often come from several sources and may contain differing values of activity. For each ligand/target complex studied, we did not possess activity values for those ligands not known to be active to the target, and hence our inactive or decoy set. Thus, rather than treat the problem as a multi-response regression with missing data, or mixed categorical and real valued data, the analysis was performed as a classification model by representing compounds as active if there was any response above an active threshold, and compounds not known to be active as inactive compounds.

Much like clustering, for predictive modeling (supervised learning), it is best to run a battery of several cross-validated models to arrive at a determination as to which model (or models) is best. For this study, with just a few variables, and using a simple two-class model (actives and inactives) we chose *K*-nearest neighbor (KNN), a classification tree algorithm, Fisher's linear discriminant, and quadratic discriminant methods to build predictive models.

The ligand/target complexes that are studied here are not data rich and so it is difficult to use the common training, validation, and testing set divisions ordinarily created for model performance. In these studies, we used *k*-fold cross-validation with simply a training set and a testing set in roughly 3/5th, 2/5th portions respectively, where the testing set sample is randomly chosen from one third of the total testing set for each run with replacement. This method helps facilitate modeling when there is a paucity of the actives vs. inactives in the data set. Hundred-fold cross-validation was performed on each model largely to better sample the wide variance of the enrichments.

Results

Cox2

In Figure 6, 43 highly active Cox2 inhibitors, as reported by the WD [26] are compared to the X-ray structure for a known Cox2 inhibitor, SC-558, PDB ID 6COX [25]. The figure displays the

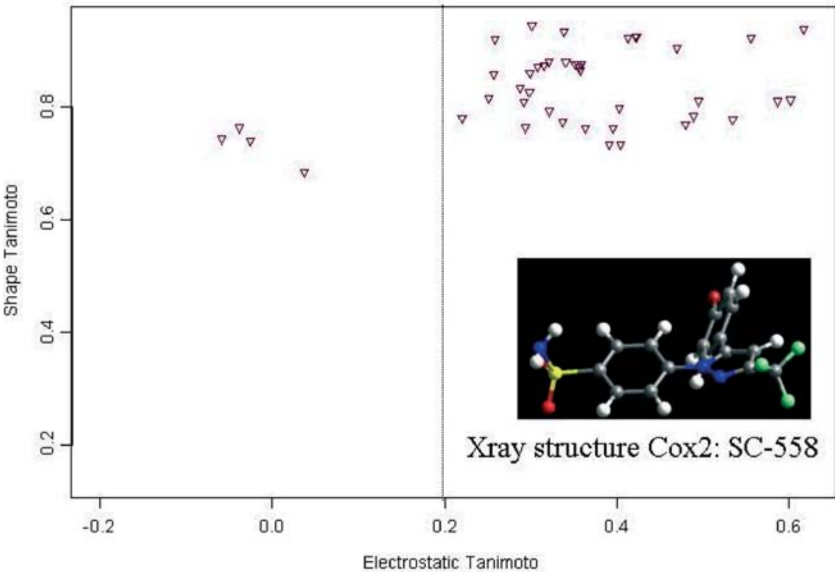

Figure 6. Shape and Electrostatic Tanimoto similarities to Cox crystal ligand SC-558 for 43 Cox2 highly active (activity greater than 10 nM) structures from the WD.

Shape Tanimoto vs. the Electrostatic Tanimoto, one conformer per SMILES, determined by the geometric mean method described above. All but four SMILES have Electrostatic Tanimoto scores above 0.2, and all but one SMILES have Shape Tanimoto values greater than 0.7. As a validation of the shape and electrostatic 'signal' observed in Figure 6, 1781 decoys (all active ligands to other targets, but not known to be active to Cox2) from the WD vs. SC-558 were similarly analyzed and the results are displayed in Figure 7. In Figure 7, one can see a distinct separation in shape and electrostatic space for all but four of the highly active Cox2 inhibitors and the decoy set. Less than one tenth of 1% of the decoy set have Shape Tanimoto scores greater than 0.7, with only one of

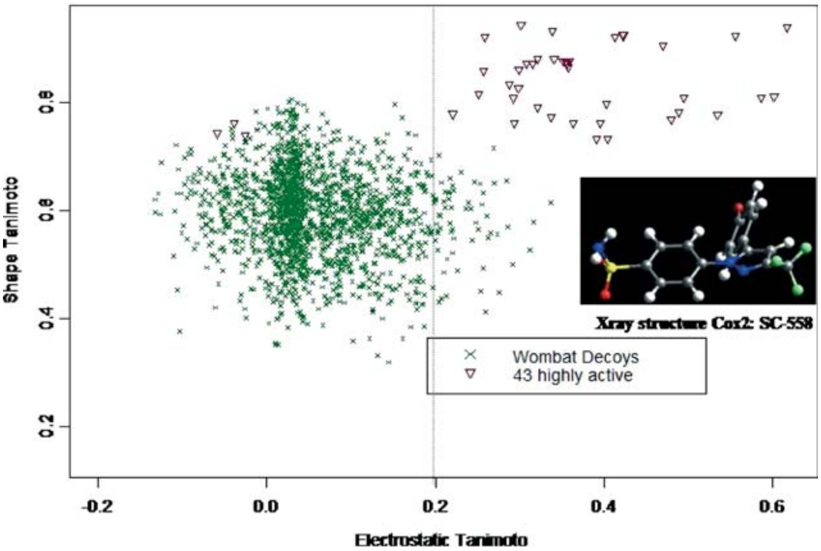

Figure 7. Decoys with no reported activity for Cox2 and 43 highly active ligands (activity> 10 nM) for Cox2 compared with SC-558, Shape Tanimoto and Electrostatic Tanimoto comparison.

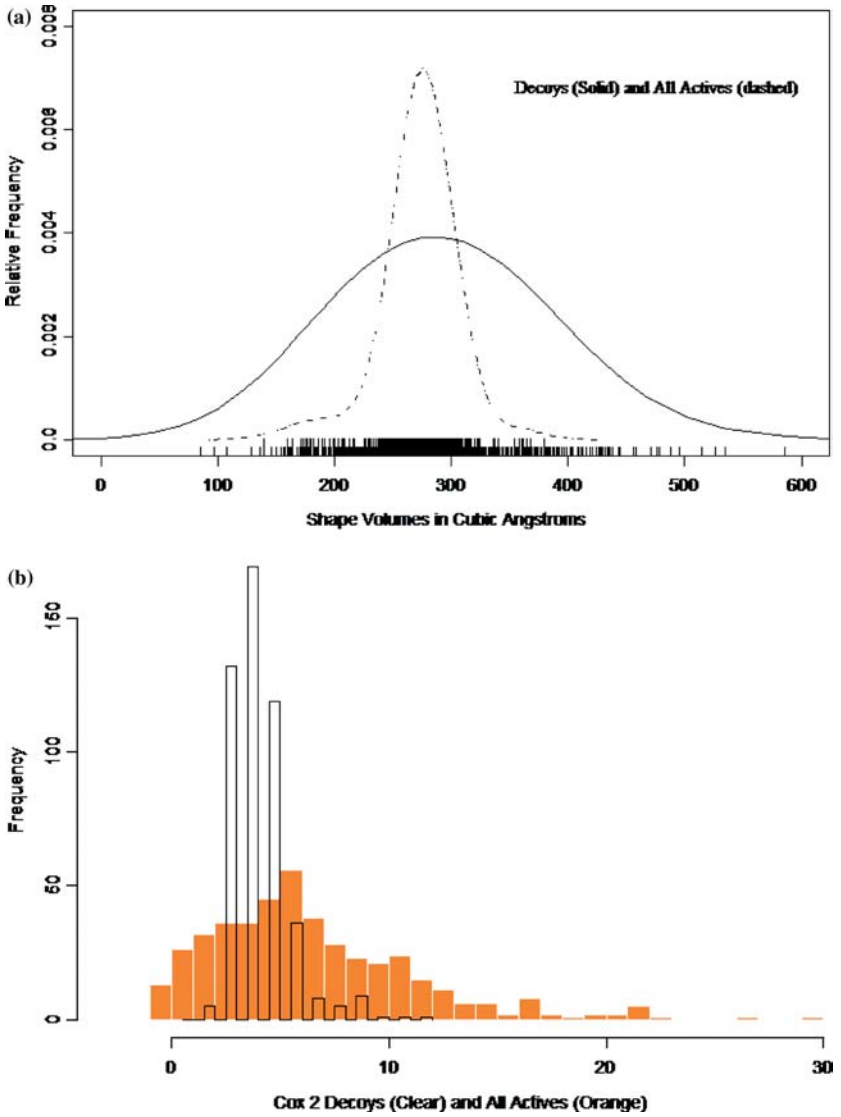

Figure 8. (a, top) The distribution of the volumes of the entire active set is fully within the distribution of the decoy set. The decoy set has a wider distribution but the modes are nearly the same for both distributions. (b, bottom) The distributions of the number of rotatable bonds. Again, the modes are quite similar, but the active set distribution is contained within the range of the decoy distribution.

those having an Electrostatic Tanimoto score greater than 0.2. In Figure 8a and b, the distribution of the decoy volumes and the decoy number of rotatable bonds suggests that these variables are not biasing the separation between the actives and the decoys found with shape and electrostatic variables. A random subset of the entire decoy set was chosen to be roughly equivalent in size to the entire active set, so that the density estimation of the volume distributions and the rotor histograms would have the same scaling.

In Figure 9 the moderately active (285 structures) and weakly active (162) ligands towards Cox2 were also included in the analysis. One still sees a separation in shape and electrostatic space for the active ligands vs. the decoy, however, the separation is weaker the less active the set. One also may notice a set of actives, boxed in Figure 9, with high shape similarity to SC-558, but low electrostatic affinity. We shall refer to this set as the Cox2 anomalous group. To graphically visualize the substantial class separation, a simple

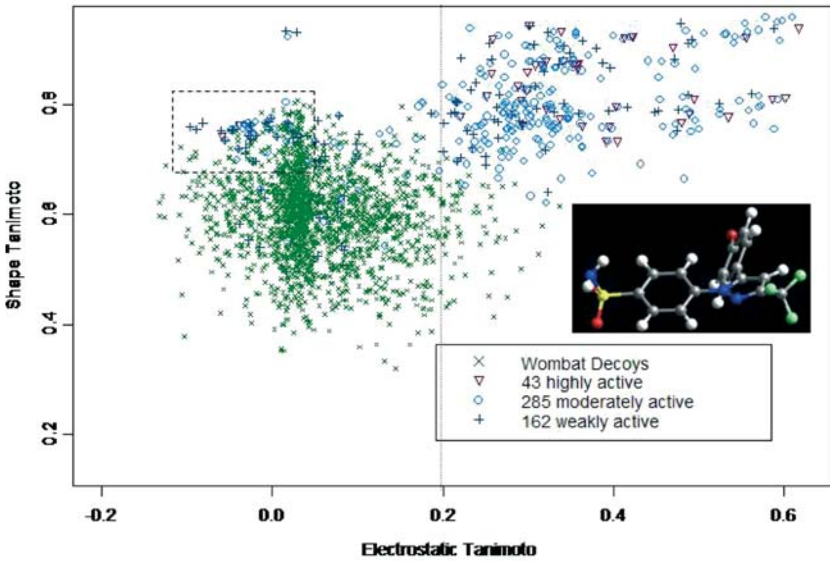

Figure 9. Wombat decoys, 43 highly active (activity > 10 nM), 285 moderately active (1 μM ≤ activity ≤ 10 nM), and 162 weakly active (0 < activity < 1 μM) ligands towards Cox2 vs. SC-558, Shape Tanimoto vs. Electrostatic Tanimoto. Cox anomalies are in the dotted box. When the anomalous region is removed, the geo-mean error increases to 0.95.

Fischer's Linear Discriminant decision boundary is displayed in Figure 10, helping to show the classification error between all active ligands towards Cox2 and decoys towards Cox2. In this simple model the accuracy, as defined by the geo-mean2 normalization, is 0.81 with a standard deviation of 0.03. If the dashed box of anomalies and the few decoys therein in Figure 9 are removed, the geo-mean2 accuracy increases to 0.93.

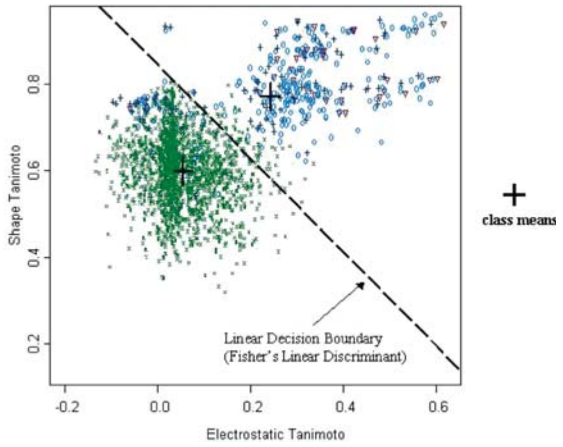

Figure 10. Fischer's linear discriminant model for the classification error between active ligands towards Cox2 and decoys towards Cox2. The large crosses represent the class means for each category (active vs. inactive). Classification error is 0.81 and standard of error is 0.03.

Table 1 shows results of the use of the Fisher linear discriminant with all combinations of Shape Tanimoto, Electrostatic Tanimoto, and the 2D Tanimoto with the MDL 320 fingerprints. Results, with and without the anomalies, are reported in terms of enrichments, along with the geo-mean2 accuracy. In general, the combination of all three variables provides the largest enrichments but not necessarily the largest accuracy. Enrichments were calculated naturally on the basis of the prior ratio of the training set, whereas the accuracy was calculated as is customary with the testing set sample ratios per cross-validation run. This created a slight discrepancy between the increasing enrichments per addition of variables and the corresponding accuracy values. Additionally, the effect of the increasing standard deviation, and therefore the variance, with the addition of variables of the *k*-fold cross-validation found in the table was addressed in the *Statistical learning theory: similarity searching* portion of the Theory section. Thus, overall the results are largely consistent. It is important to note, however, that when the anomalies are removed, the enrichments drop substantially. This is due to the fact that now the separation with the anomalies removed is considerably better, and getting close to near perfect separation, leaving little room for large

Table 1: A simple Fisher linear discriminant was performed on the Cox2 data with decoys and all actives, both with and without anomalies.

	Data	Number	Percent		
	Fisher's linear discriminant				
	Actives (high, moderate, and weak)	487	21.47		
	Inactive decoys	1781	78.53		
	Variables	Enrichments		Geo-mean2 error	
		Mean	STD	Mean	STD
1D	2D Tanimoto 320 MACCS keys	9.05X	2.23	0.76	0.03
1D	Shape Tanimoto	7.96X	1.53	0.87	0.02
1D	Electrostatic Tanimoto	6.2X	1.32	0.76	0.03
2D	Shape and Electrostatic Tanimoto	26.31X	13.03	0.81	0.02
2D	320 and Electrostatic Tanimoto	13.28X	4.06	0.79	0.03
2D	320 and Shape Tanimoto	21.84X	6.45	0.88	0.03
3D	All three	33.94X	13.21	0.83	0.03
	Data	Number	Percent		
	With anomalies removed				
	Actives (high, moderate, and weak)	399	17.48		
	Inactive decoys	1683	82.52		
	Variables	Enrichments		Geo-mean2 error	
		Mean	STD	Mean	STD
1D	2D Tanimoto 320 MACCS keys	3.84X	0.6	0.85	0.01
1D	Shape Tanimoto	7.25X	1.34	0.92	0.01
1D	Electrostatic Tanimoto	6.50X	1.14	0.91	0.01
2D	Shape and Electrostatic Tanimoto	8.75X	1.68	0.93	0.01
2D	320 and Electrostatic Tanimoto	8.11X	1.78	0.93	0.01
2D	320 and Shape Tanimoto	8.76X	1.87	0.93	0.01
3D	All three	11.5X	2.25	0.95	0.01

Results are reported in terms of enrichments and geo-mean2 accuracy. Active ligand classes used: highly active (activity > 10 nM), moderately active (1 μM ≤ activity ≤ 10 nM), weakly active ($0 < \text{activity} < 1$ μM).

enrichment values between the prior ratio and the decision boundary determined ratio.

Three other classification models were also performed with all actives with decoys and all three variables: KNN, tree classification, and a quadratic discriminant. Of the four models, the quadratic discriminant performed the best with a geo-mean2 accuracy of 0.88. KNN, linear discriminant, and the tree classification model accuracy values were 0.86, 0.83, and 0.65, respectively. All had relatively small standard deviations of 0.03 or less. With this data, in general, the linear and quadratic discriminants, and KNN were roughly equivalent, whereas the tree classifier performed significantly worse. It is clear via visual inspection of Figure 10, and the model performance accuracy values of the three best models, and the enrichment value that the actives can be well separated from the inactive decoy set.

The anomalous group described above provides an excellent example of 'negative' information. These anomalies are only modestly shape similar, and not at all electrostatically similar to SC-558, so the question arises: do they represent another binding mode or binding site? A shape 'centroid' molecule was selected from the anomalous group using Shape Tanimoto within the group. The shape and electrostatic comparison to the remaining anomalies is displayed in Figure 11a and b. Figure 11b also displays the 2D diversity of the anomalous group, determined by the use of the MDL 320 fingerprints and the Tanimoto similarity

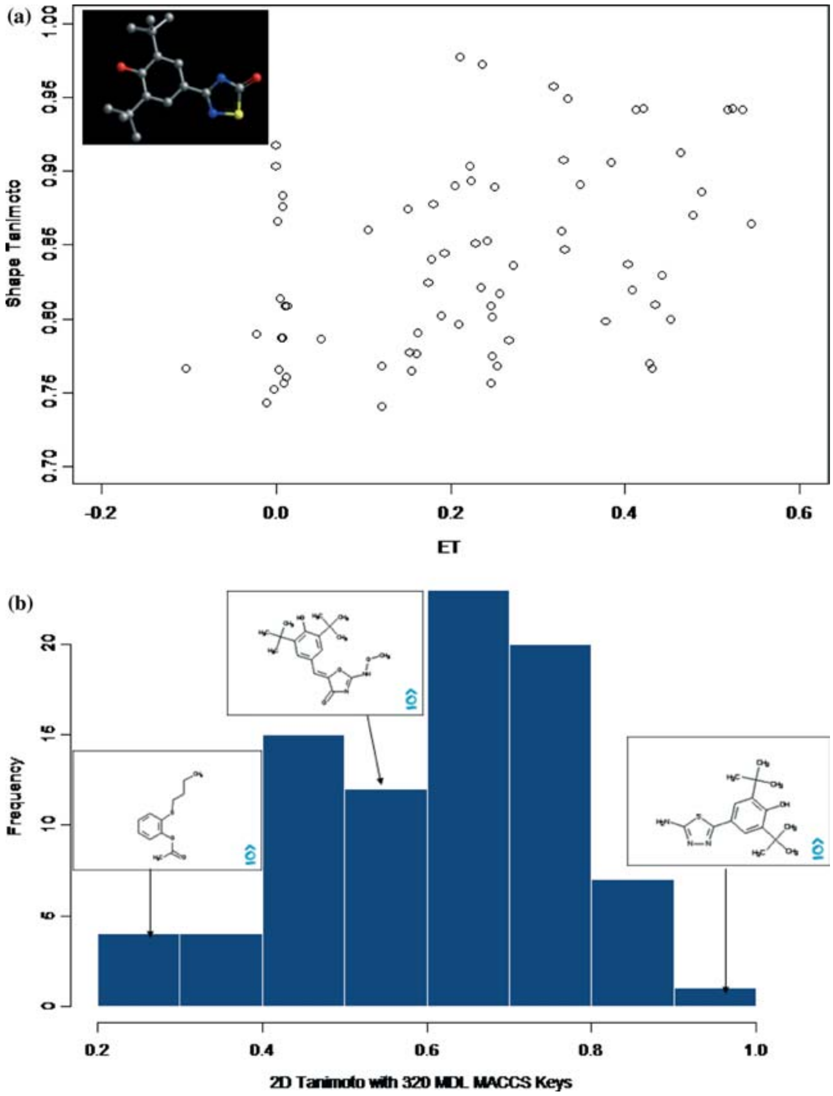

Figure 11. (a, top) Cox2 anomalous ligands vs. anomalous centroid 2D structure analysis. Shape Tanimoto vs. Electrostatic Tanimoto of the centroid vs. the anomalous ligands is given in the plot. The centroid structure is depicted in (a). (b, bottom) The histogram represents the distribution of 2D Tanimoto values between the anomalous ligands and the centroid. Example 2D anomalous ligands are displayed for several of the histogram bins.

coefficient. All anomalies have a Shape Tanimoto of 0.70 or greater to the centroid, while their 2D Tanimoto scores range from 0.2 to 0.9. The majority of the electrostatic scores in Figure 11a are greater than 0.1.

Progesterone

In Figure 12 the Shape and Electrostatic Tanimoto values between the X-ray structure of progesterone (PDB code 1A28 [25]) and weakly active, moderately active and highly active ligands as well as 1000 decoys from the WD are displayed. Decoys were randomly selected from a subset of the WD in which all known ligands of progesterone were removed. Contrary to the nice separation that we observed for the Cox2 analysis, progesterone actives and decoys are highly overlapped. Figure 13 shows the progesterone receptor with bound, and completely enclosed, progesterone. A secondary proximal pocket is displayed. The depth and tightness of fit of the progesterone pocket, as well as knowledge that

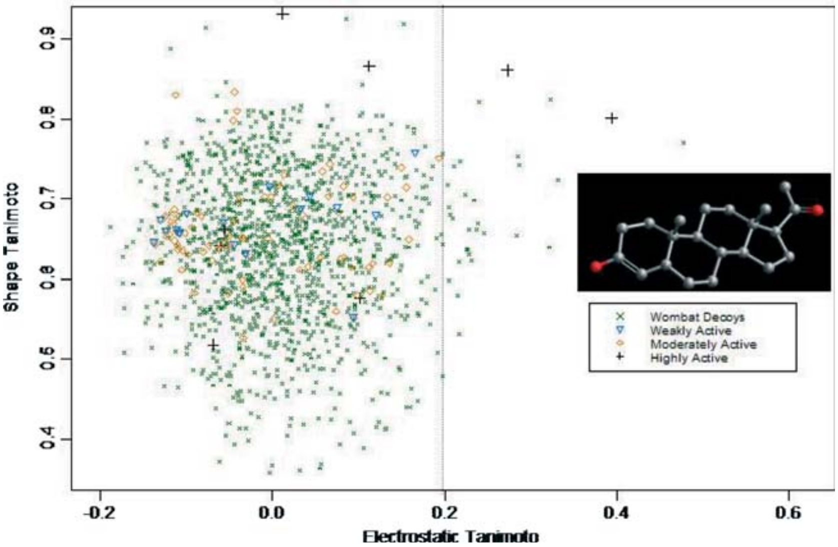

Figure 12. Comparison of Shape Tanimoto vs. Electrostatic Tanimoto between the X-ray structure of progesterone and WD decoys, weakly active ($0 < \text{activity} < 1\ \mu\text{M}$), moderately active ($1\ \mu\text{M} \leq \text{activity} \leq 10$ nM) and highly active ligands ($\text{activity} > 10$ nM).

progesterone causes an induced-fit, led us to consider if the larger actives found in the WD are actually binding to a different location or an open vs. closed form of the protein. The results for the 100 active progesterone ligands clustered on Shape Tanimoto are displayed in Figure 14. Two main clusters of conformers are quite distinct: conformers for 23 compounds are contained in Cluster A, conformers for 64 compounds in Cluster B, and conformers for the remaining 12 compounds are in three

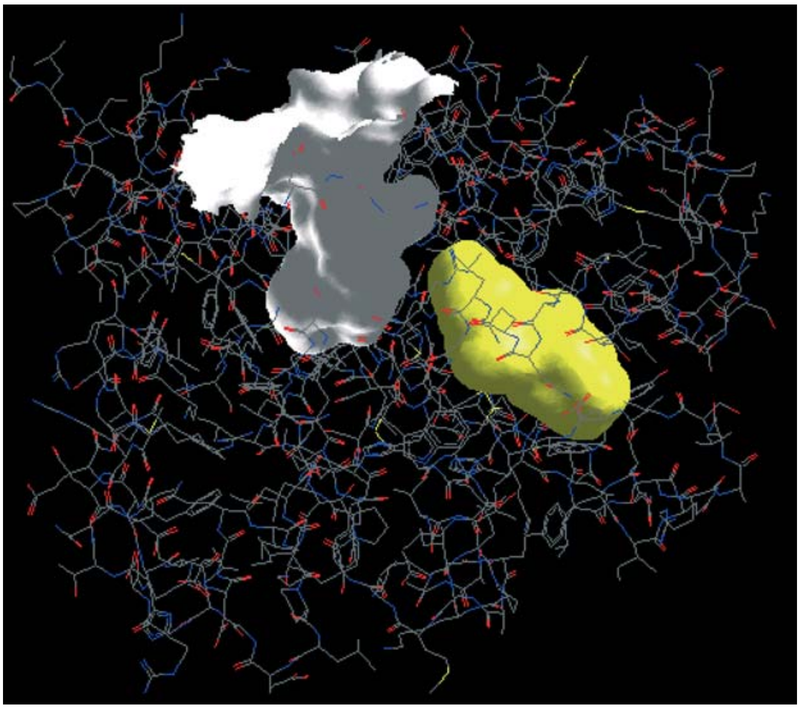

Figure 13. Progesterone receptor with bound progesterone, shape surface (yellow) and second possible binding site (white).

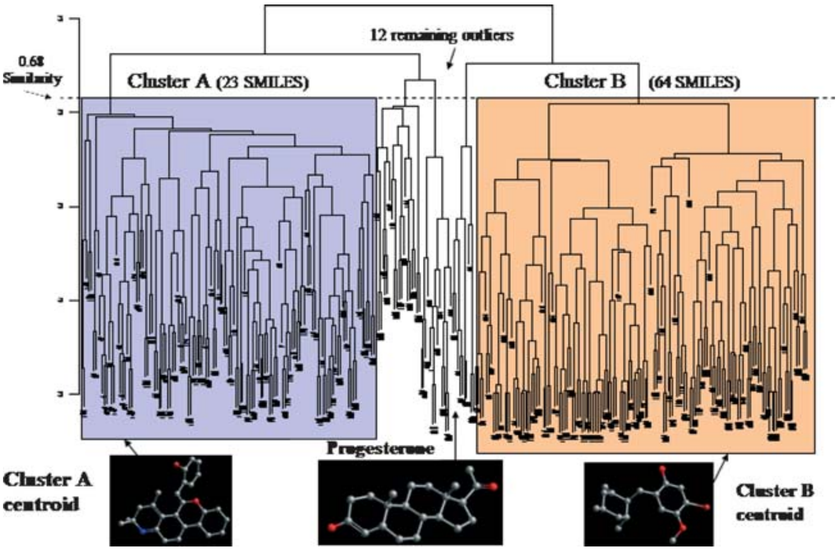

Figure 14. Group average clustering 100 progesterone active SMILES. Cut made at 0.68 similarity to pick up ~25% of SMILES in Cluster A and ~65% of SMILES in Cluster B.

small outlier clusters. Progesterone conformers are contained in one of these outlier groups. The centroid molecular structure for Cluster A and Cluster B was determined and used as a target against the remaining members of each cluster. The results for Cluster A are given in Figure 15 and the results for Cluster B are given in Figure 16. The Cluster A vs. the Cluster A centroid separate nicely in shape and electrostatic space from the decoy set, with all shape scores in Figure 15a greater than 0.75 and all Electrostatic Tanimoto scores greater than 0.2. The structural diversity of Cluster A is displayed in the 2D Tanimoto vs. Electrostatic Tanimoto and the 2D Tanimoto vs. Shape Tanimoto plots, Figures 15b and c. The cluster has many series, not just that of the Cluster A centroid. The results for Cluster B in Figure 16a show 13 of the 64 cluster members with Shape Tanimoto scores greater than 0.75 and Electrostatic Tanimoto scores greater than 0.2. The 2D Tanimoto scores given in Figure 16b and c are all greater than 0.75, so Cluster B contains SMILES from similar structure series. The centroid from Cluster B is the shape centroid for the cluster.

Cluster B does not show such clear separation of actives from decoys as Cluster A. This could be because the shape centroid method does not simultaneously select an electrostatic centroid, i.e. the centroid may be a poor representative of the electrostatic profile of this set of actives. However, the separation is far from poor and suffers more in comparison to Cluster A than in absolute terms.

Dopamine

In this experiment there was no 'target' structure. A set of Dopamine active ligands from the WD were shape-clustered and three groups were found. The shape centroid from each was compared with its cluster members and with a set of a 1000 WD decoys. The WD decoys were randomly selected from a subset of the WD database in which ligands with known Dopamine activity were removed. The Shape vs. Electrostatic Tanimoto values are displayed in Figure 17. The first centroid is able to recover all but three of its cluster members, with Shape Tanimoto scores greater than 0.7 and Electrostatic Tanimoto values greater than 0.8, shown in Figure 17a. Figure 17b shows that the 2D Tanimoto values for this cluster range from 0.45 to 0.9. This cluster represents a range of 2D structural diversity, while maintaining a shape and electrostatic 'signal'. The same clear separation in shape and electrostatic space is observed for the remaining two cluster centroids and their members, as displayed in Figure 17c and d.

Calcium ion channel

The same shape clustering analysis applied to the Dopamine ligand experiment was applied to a set of calcium ion channel ligands. A separation in shape clustering was not discovered at reasonable Shape Tanimoto thresholds. This was not

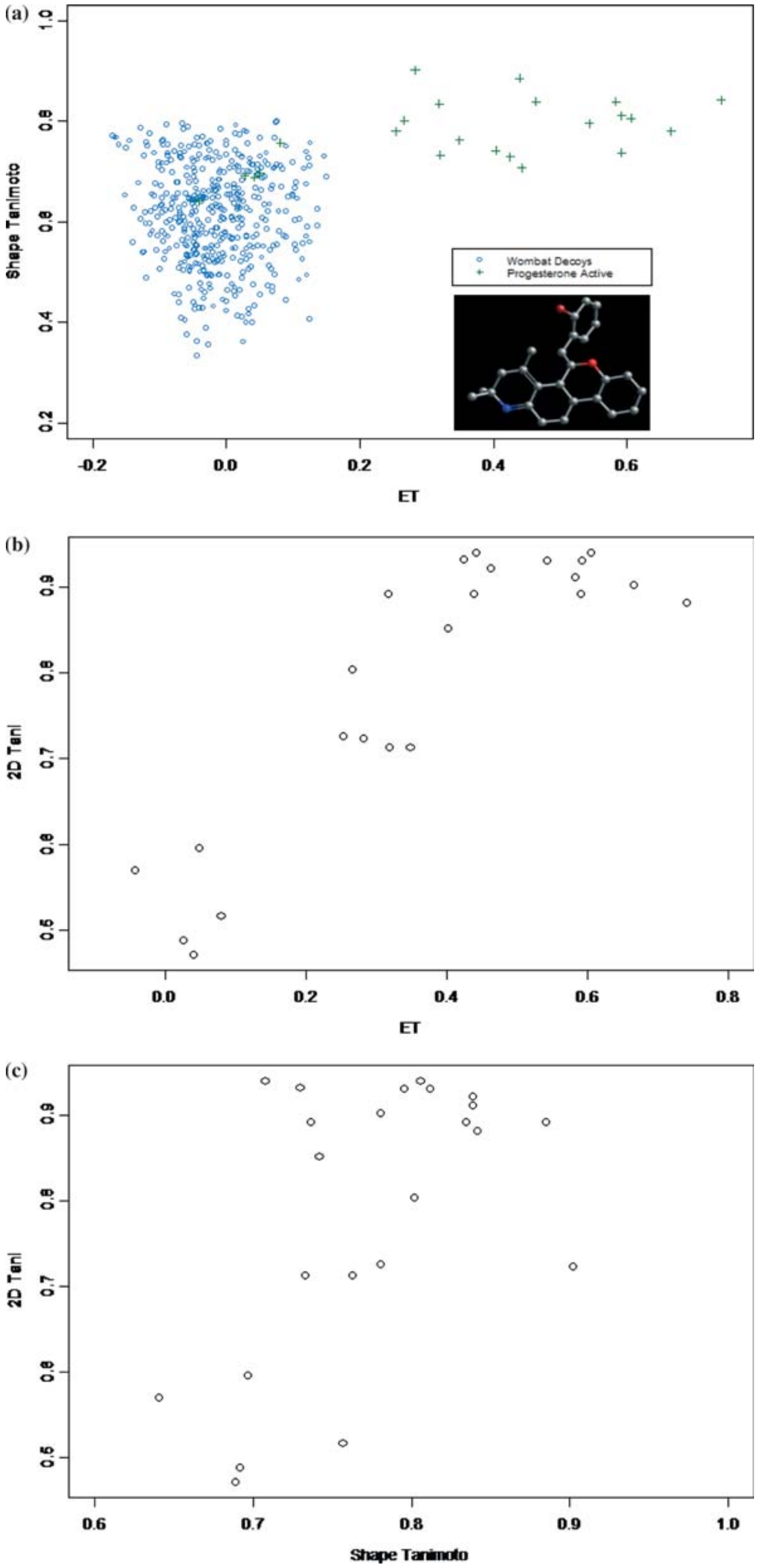

Figure 15. (a) The Shape Tanimoto vs. Electrostatic Tanimoto (ET) comparisons of Cluster A members and decoys vs. the centroid from Cluster A. (b) Similarly for the 2D Tanimoto vs. Electrostatic Tanimoto. (c) The 2D Tanimoto vs. the Shape Tanimoto.

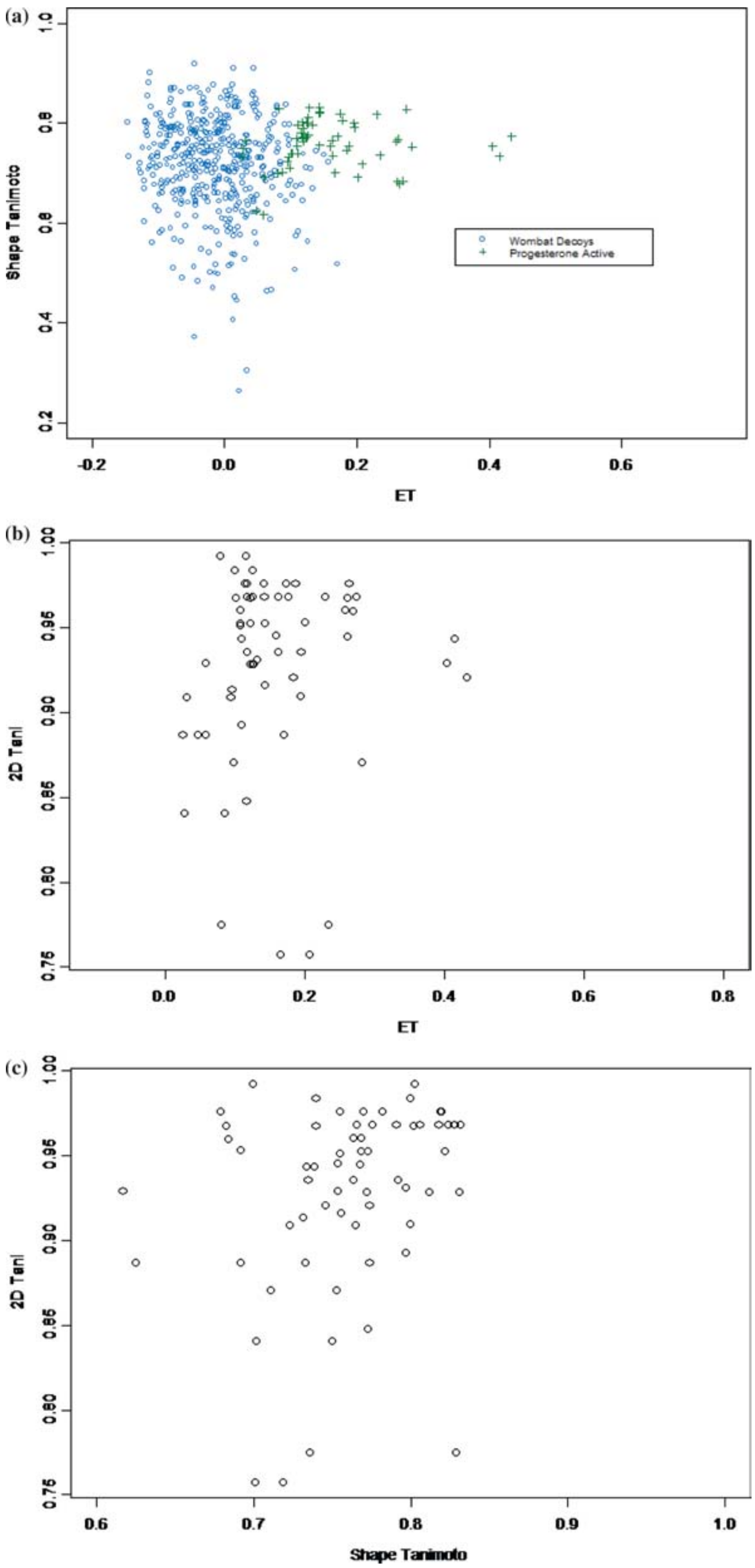

Figure 16. (a) The Shape Tanimoto vs. Electrostatic Tanimoto (ET) of Cluster B members and decoys vs. the centroid from Cluster B. Similarly, (b) displays the 2D Tanimoto vs. Electrostatic Tanimoto (ET), and (c) displays the 2D Tanimoto vs. the Shape Tanimoto.

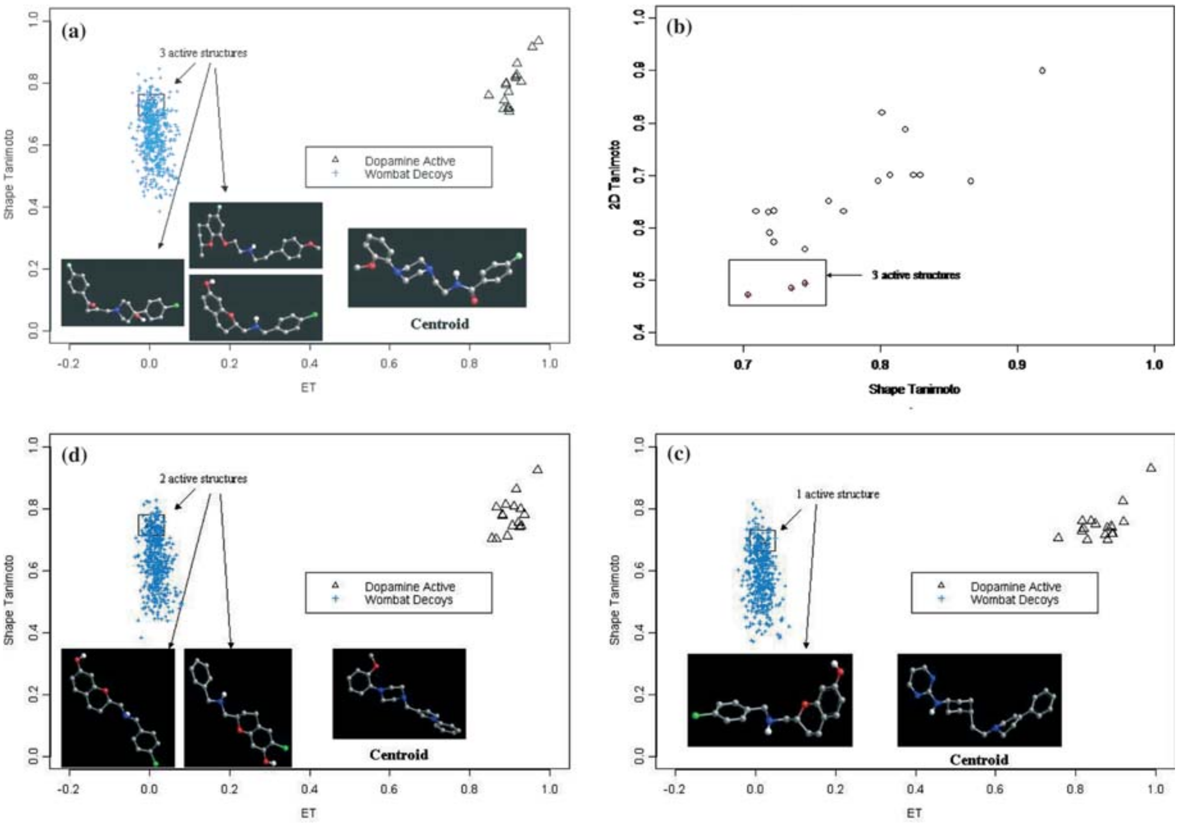

Figure 17. (a–d) Shape Tanimoto vs. Electrostatic Tanimoto scores for three Dopamine cluster centroids vs. cluster members.

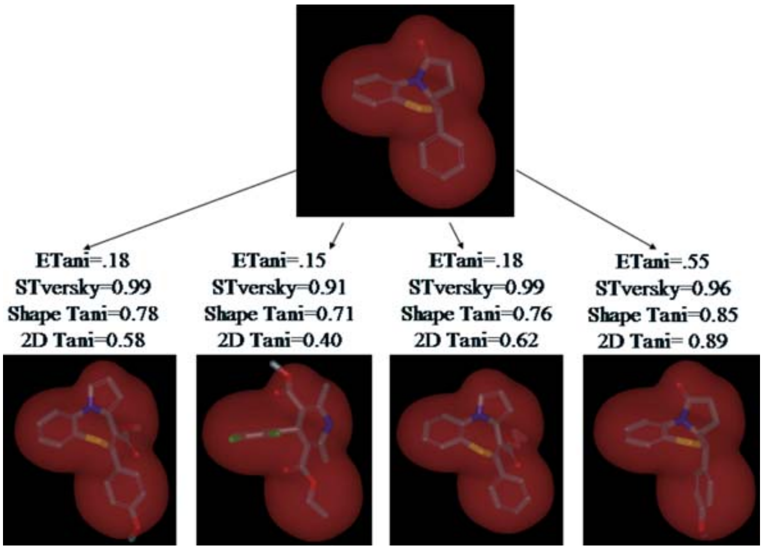

Figure 18. The cluster centroid at the top of the figure is compared to four other cluster members. The asymmetric TB grouping algorithm found four electrostatically similar Ca ion channel ligands. Comparisons of Electrostatic Tanimoto, Shape Tversky, Shape Tanimoto and 2D Tanimoto are displayed.

surprising given the diversity in size of the ligands. We then clustered by sub-shape using the Shape Tversky measure and the asymmetric TB leader algorithm. One 'sub-shape' was found to be common amongst 5 of the 12 ligands analyzed. Figure 18 displays the overlays and scores in Electrostatic Tanimoto, Shape Tversky, Shape Tanimoto, and 2D Tanimoto between the 'sub-shape' and the four electrostatically similar calcium ion channel ligands. A clear pattern in 3D shape and electrostatic space has been found starting from just SMILES and activities.

Discussion

The results presented here are the first attempts at applying a rigorous, formal, and general theory of shape to molecular activity prediction. The term 'shape' has been used often in the field of molecular comparison, meaning many things, often unrelated to three-dimensional properties. However, we justify our use of the word, and our claim that this is the first application of such, by three precepts. The first is that the description of shape is precise, mathematical, and complete. There are no arbitrary parameters, no conditions on types of molecule or types of chemistry. It is a formal theory because it reduces a common conception to a precise, unambiguous, form. Other uses of the words 'molecular shape' may also be precise, but not general. Shape has, unfortunately, been applied to almost anything that has the mere semblance of its common usage. Secondly, this definition of shape, unlike others, produces a metric quantity, and therefore one that enforces a structure on the distribution of molecules in a 'shape space'. Other work [12] examines this property more thoroughly, but, in our work, it means that structures that are similar to each other are not going to also be similar to other structures, unless they all are – a basic application of the triangle inequality. We believe this is reflected in the outstanding separation of actives in Cox2, progesterone, and, especially, Dopamine. Thirdly, our definition of shape is not restricted to chemistry. An analogy here is with PB theory as applied to molecules. PB theory is completely general; it can be applied to molecules but also to large systems. This can be seen in examples such as transistor circuits or capacitor design. It has no scale-dependence, except possibly at the atomic level where the concept of a local dielectric might fail (but seems not to). Similarly, our definition of shape can be applied to any three-dimensional (or higher or lower dimensional) body. It could equally well be used to organize Grecian urns as molecules. Some would see this as a disadvantage, in that it is lacking the local domain knowledge of chemistry. However, it can be argued that this is the problem with many descriptor methods: that they encode chemical perspectives, not fundamental properties. Finally, we note that we could include electrostatics here as a type of shape, given that the form of comparison is identical, although in this paper it acts as a dependent variable of steric shape.

An obvious criticism of shape herein is that one may be able to define such quantities but are they relevant to the problems at hand? Molecular weight similarity is a well-defined concept but it does not help (much) in drug design, other than as a filtering property to avoid problems in bioavailability, solubility or synthetic cost. Our response is several-fold. Firstly, what we have done is codified common experience, i.e. that molecules that look the same tend to act the same. This lies behind much of our intuition of molecular behavior. For instance, modelers know that high-affinity ligands to proteins always fit well, i.e. the lock and key hypothesis of Fischer, proposed in 1890 [32], is still dominant because it agrees with observation. (Docking is an instance of shape complementarity, where here we have dealt with shape similarity.) In this sense the method 'has to work' to some degree. The second comes from the striking empirical evidence. We did not choose the four examples in this paper because of their success; we chose them as the first four test cases we have examined. In addition, the tests have been of increasing difficulty: from a well-defined target and active site, to an active site requiring induced-fit, to assay data alone with no ligand structure. That the results are clear-cut as they are is a testament to the method, not judicious application. We are currently conducting tests on an equal number of additional experiments with data provided by industrial partners and anticipate that number growing greatly when the procedure can be made readily available. Third, is the manner of failure. Hypotheses based on good theory fail gracefully, and shape here fails well. If we are asking the wrong question, we should expect a definitive 'No' and this is exactly what is delivered. In the case of progesterone, some actives studied bind to a different site. With Cox2, the anomalies may represent either a different binding mode or a fragment of the same binding mode. And, in the calcium channel blockers, where the test set is too diverse in shape to represent a global solution, a local solution can be inferred. (Lack of locality in the theory shall be discussed below.) Thus, the power of negative information appears to be just as useful as the positive information. Finally, a case can certainly be made that molecular interactions are so dominated by shape and electrostatics that any method that encapsulates such must reflect the realities of biological efficacy. Van der Waals interactions and hydrophobicity are shape-dominated, hydrogen bonds, π–π interactions (e.g., stacked rings), polar/charge interactions are all dominated by electrostatics. This is not to say that there are not other potential effects (most notably, conformational entropy and induced polarization/ionization) but that shape and electrostatics clearly capture the largest fraction of molecular interactions. Other approaches, for instance, 3D pharmacophores, or atom-typing in chemical fingerprints, do so more remotely via classification theory.

It is worthwhile to consider why shape has worked well here, given the perception of 3D methods is that they are not powerful without in-depth information. We believe that the primary reason is attention to the details necessary to work with 3D structures. Some approaches have settled for single conformers representing each molecule (missing information) or, at the other extreme, have used large ensembles without any energy weighting (overwhelming any signal from the most likely conformers). Other approaches have used inappropriate charging schemes, whereas we have attempted to use charges that reflect physical observables (dipole moments, vacuum–water transfer energies, electrostatic potential from *ab initio* calculations). In addition we have used a method that will produce potentials as seen in solvent, not in vacuum. It has been shown that this will improve other 3D methods, e.g., CoMFA [33], just as having better atom-based charges improves physical property prediction. The use of the ROCS procedure has clearly been important in that it attempts to find the best global alignment of structures, not just that from some clique of pharmacophores, or by moment of inertia, or by alignment of common substructures. The fact that ROCS aligns pairs of molecules, not ensembles, may be seen as a disadvantage. However, the fundamental property of shape is defined here as a pairwise property and as such ROCS is a reliable tool. Finally, we believe many approaches have been hampered by incorrect or incomplete chemical informatics. All the tools here were built on one CIS, that provided by OEChem, which we believe offers the most reliable foundation available. Our general conclusion is that the more accurate the physics applied to molecules, better charges, better structures, etc., the better the results. We have seen this in property prediction and docking and believe the same principle applies to QSAR. The determination of activity is non-trivial or the pharmaceutical industry would not spend billions of dollars a year on research. Hindering the process with less than complete descriptions of molecules does not make it any easier.

There are many potential criticisms of our approach. Although the metric properties of shape and electrostatics distances are fundamental, there is no reason that the correlation with activity be a strong one. Our work suggests the correlation is substantial but that may be disproved by future examples. In addition, we have only demonstrated classification, not activity prediction, as such regression methods as CoMFA attempt. However, the Cox2 results suggest that such a theory could be derived and this will be a future direction of the work. A more substantial criticism is that our similarity measures are global and hence can fail to predict correlations that are based on local features. A drug binding to an exposed active site may make interactions with only a part of its shape. In a similar vein, it is clear that local chemistry is important for some processes, such as p450 activity with consequences for toxicity. Chemical activity may be more important than the physical properties of shape and electrostatic profile. It is possible therein lies the substantial synergy with 2D fingerprints – such limitations do not apply when one is using chemical connectivity, which is essentially a local property. The same could be said of a clique pattern of pharmacophores. This can be ameliorated, to some extent, with Shape Tversky, and this approach could be extended to electrostatic partial similarity, but this has yet to be demonstrated.

Another criticism is novelty: we are not the first to use the comparison of molecular fields and we are not the first to use Gaussians to represent shape [34–36]. However, the difference in our approach is substantial: we use Gaussians only because we have shown that they can be used to represent a molecule's characteristic (volume) function, not as an extension of a pharmacophore point for steric effects [34–36]. Our electrostatic field comparison is predicated on prior shape comparison, which is pairwise, not groupwise (as in CoMFA). We do not use any chemical information or patterns for shape or for electrostatic representations. Our choice of metric property is unique and different from the index used by some for the comparison of fields from quantum calculations (the 'Carbo Index'). Our method does not require any hypothesis generation (i.e. divining the appropriate pharmacophoric groups) or parameterization. We imagined, initially, that the method would require a good 'target', one that would have a known structure from a protein–ligand complex (Cox2) or a rigid chemical structure (progesterone), but this has turned out not to be the case: shape and electrostatic patterns can be derived from activities alone via shape or partial shape clustering. There needs to be more evidence that this approach will work reliably, but it is possible that the strong metric properties of shape make this a feasible, general approach.

Future work and conclusions

Clearly the combination of shape and electrostatics as introduced in this paper needs to be more widely tested, both by us and by others in the field. All the steps in the process described can be duplicated either with the tools described or variants thereof. We anticipate organizing the different software elements required for the work presented here into a single program to facilitate widespread use. In addition, the development of simple regression models for activity will be examined in the hope that more information can be derived as to relative activities. Our comments above concerning physical accuracy in models have not been fully tested. We expect the results to reliably degrade with poorer physical theories and improve with better theories, but that can now be examined thoroughly. In the case of atom charges we can improve our descriptions with the Bayly method [37]. Conformer energies we can improve with the addition of implicit solvent forces, semi-empirical or *ab initio* energies.

One improvement already investigated is a variant of EON that allows some manipulation of the target or query structure. Atoms belonging to small terminal rotors that have significant charge can affect target–query ET without significantly impacting the ST, the Shape Tanimoto. EON-SPIN performs a rotor search for these atoms, recalculating the ST and ET, reporting the pair that has the best ET. One reason this is necessary is that structure generation typically uses the RMS difference between coordinates of different conformers to reduce the number of representative structures. However, small terminal rotors make very little difference to the RMS between structures and hence are 'under-represented' in typical conformer collections. EON-SPIN essentially 'rescues' such conformations if they have significant electrostatic similarity.

The generalization of shape matching to local properties can be explored within this framework, in particular, electrostatics which has more local behavior than steric properties: it is easier to vary a local potential than a local shape. Our test case of calcium ion channel blockers is one example. It now appears possible to extend the work to the discrimination of agonists and antagonists [38]. Such data is available from the WD but has not yet been included in our analysis.

It remains to be seen if the signals for activity by shape and electrostatic field comparisons will be as strong for other types of phenomena, e.g. toxicity, bioavailability, and active-transport. We anticipate there may be applications in these areas, but not all. Clearly our *ansatz* is rooted in the physics of molecular interaction: membrane permeability may not correlate at all with molecular shape. However, we have been surprised with the method's success so far and look forward to its more widespread examination.

In conclusion, we have demonstrated a simple framework with which the concepts of molecular shape and electrostatics can be applied to problems of interest in the QSAR paradigm. The theory is concise, general and appears to work.

Acknowledgements

The authors wish to thank Tudor Oprea of Sunset Molecular Discovery, LLC for the use of the Wombat Database; and Geoff Skillman, Roger Sayle, Jeremy Yang, and Robert Tolbert of OpenEye Scientific Software, Inc., for their helpful discussions, and software assistance and support.

References

1. Daylight Theory Manual, Daylight CIS Inc., Mission Viejo, CA, http://www.daylight.com.
2. Barnard Chemical Information Ltd., Sheffield, UK, http://www.bci.gb.com.
3. MDL Information Systems, Inc., San Leandro, CA, http://www.mdli.com.
4. BioByte, Claremont, CA, http://www.biobyte.com.
5. Edusoft, San Francisco, CA, http://www.edusoft.com.
6. Cramer, R.D., Poss, M.A., Hermsmeier, M.A., Caulfield, T.J., Kowala, M.C. and Valentine, M.T., J. Med. Chem., 42 (1999) 3919.
7. Pearlman, R.S. and Smith, K.M., J. Chem. Inf. Comput. Sci., 39 (1999) 28.
8. Greco, G., Novellino, E. and Martin, Y.C., In Lipkowitz, K.B. and Boyd, D.B. (Eds.), Reviews in Computational Chemistry, VCH Publishers, New York, NY, 1997, pp. 183–240.
9. Wang, R., Fu, Y. and Lai, L., J. Chem. Inf. Comput. Sci., 37 (1997) 615.
10. Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J.G., J. Chem. Inf. Comput. Sci., 42 (2002) 1273.
11. Masek, B.B., Merchant, A. and Mattheews, J.B., Proteins, 17 (1993) 193.
12. Gower, J.C. and Legendre, P., J. Classif., 3 (1986) 5.
13. Halgren, T.A., J. Comput. Chem., 17 (1996) 490.
14. Halgren, T.A., J. Comput. Chem., 20 (1999) 720.
15. Gasteiger, J. and Marsili, M., Tetrahedron Lett., (1978) 3181.
16. Fingerprint Module, Mesa Analytics & Computing, LLC, Santa Fe, NM, http://www.mesaac.com.
17. OEChem – C++ Theory Manual, OpenEye Scientific Software, Santa Fe, NM, http://www.eyesopen.com.
18. Leach, A.R. and Gillet, V.J., An Introduction to Chemoinformatics. Kluwer, Boston, MA, 2003.
19. Kubat, M., Holte, R.C. and Matwin, S., Mach. Learn., 30 (1998) 195.
20. Jain, A.K. and Dubes, R.C., Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, 1988.
21. Grant, J.A. and Pickup, B.T., J. Comput. Chem., 17 (1996) 1653.
22. ROCS, OpenEye Scientific Software, Santa Fe, NM, http://www.eyesopen.com.
23. Spellmeyer, D.C., Wong, A.K. and Bower, M.J., J. Mol. Graph. Mod., 15 (1997)18.
24. Boström, J., J. Comput.-Aided Mol. Des., 15 (2001) 1137.
25. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
26. Wombat Database, Sunset Molecular Discovery LLC, Santa Fe, NM, http://www.sunsetmolecular.com.
27. Downs, G.M., Barnard, J.M., Lipkowitz, K.B. and Boyd, D.B. (Eds.), Reviews in Computational Chemistry, Wiley–VCH, New York, NY, 2002, pp. 1–40.
28. Taylor, R., J. Chem. Inf. Comput. Sci., 35 (1995) 59.
29. Butina, D., J. Chem. Inf. Comput. Sci., 39 (1999) 747.
30. MacCuish, N.E. and MacCuish, J.D., Chemometrics and Chemoinformatics, ACS Symposium Series, in press.
31. Tarjan, R., Inf. Process. Lett., 17 (1983) 37.
32. Fischer, E., Ber. Dt. Chem. Ges., 27 (1894) 2985.
33. Kellogg, G.E., Phatak, S., Nicholls, A. and Grant, A., QSAR Comb. Sci., 22 (2003) 959.
34. Kearsley, S.K. and Smith, G.M., Tet. Comput. Met., 3 (1990) 615.
35. Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.
36. Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., 33 (1993) 112.
37. Jaklian, A., Jack, D.B. and Bayly, C., J. Comput. Chem., 23 (2002) 1623–1641.
38. Katz, A.H., Tawa, G.J., Mason, K., Gove, S. and Alvarez, J., In COMP92, 227th American Chemical Society National Meeting, Anaheim, CA, 2004.