# Reasoning About Molecular Similarity and Properties

Rahul Singh
*Department of Computer Science, San Francisco State University*
*rsingh@cs.sfsu.edu*

## Abstract

*Ascertaining the similarity amongst molecules is a fundamental problem in biology and drug discovery. Since similar molecules tend to have similar biological properties, the notion of molecular similarity plays an important role in exploration of molecular structural space, query-retrieval in molecular databases, and in structure-activity modeling. This problem is related to the issue of molecular representation. Currently, approaches with high descriptive power like 3D surface-based representations are available. However, most techniques tend to focus on 2D graph-based molecular similarity due to the complexity that accompanies reasoning with more elaborate representations. This paper addresses the problem of determining similarity when molecules are described using complex surface-based representations. It proposes an intrinsic, spherical representation that systematically maps points on a molecular surface to points on a standard coordinate system (a sphere). Molecular geometry, molecular fields, and effects due to field super-positioning can then be captured as distributions on the surface of the sphere. Molecular similarity is obtained by computing the similarity of the corresponding property distributions using a novel formulation of histogram-intersection. This method is robust to noise, obviates molecular pose-optimization, can incorporate conformational variations, and facilitates highly efficient determination of similarity. Retrieval performance, applications in structure-activity modeling of complex biological properties, and comparisons with existing research and commercial methods demonstrate the validity and effectiveness of the approach.*

## 1. Introduction

Across all biological and pharmaceutical investigations, the discovery (or development) of molecules with desired biological activity continues to be an important goal. Efforts to attain this goal are strongly driven by the notion of molecular similarity because in general similar molecules tend to behave similarly [7, 13]. In contemporary research and development, applications of the notion of molecular similarity can be observed to have broadly occurred along the following three directions:

- Bio-chemical and computational exploration of the molecular structural space consisting of known (synthesized or non-synthesized) structures.
- Development of computational structure-property models that relate variations in molecular structure to variations in molecular activity or properties
- Querying of molecular structural databases.

Specific examples of the above abound. For instance, it is well known that the biological activities of proteins depend to a large extent on their binding motifs. Consequently, different algorithms have been developed to determine similarity of 3D structural motifs in proteins using backbone fragment similarity, similarity of secondary structure elements [18], or similarity of 3D configuration of residues in space [2]. The principle of molecular similarity also underlies bio-chemical approaches like affinity tagging [22], click chemistry [11], or design of chemical probes that target protein families with similar active site chemistries [6, 14]. In the domain of small molecules typical in drug discovery, usage of molecular similarity is equally ubiquitous and has been applied in areas ranging from diversity analysis for molecular library construction, molecular docking and virtual screening [9, 13, 26, 27], and a range of investigations aimed at modeling molecular structure-property relationships where the similarity of a given a molecule with a known class (or classes) of molecules is used as an important indicator of its in-vitro or in-vivo activity. Pertinent examples of such modeling include determining "drug-likeness" of molecules [1, 16], modeling blood-brain barrier permeation [15], and modeling molecular pharmacokinetics or pharmacodynamics [19, 20, 24]. Finally, use of molecular similarity is central to information retrieval

0-7695-2194-0/04 $20.00 © 2004 **IEEE**

from structural and molecular databases such as [40, 41, 42], commonly employed in biological and pharmaceutical research and development.

In this paper, we consider the problem of determining molecular similarity. Our formulation of this problem involves specification of an entire molecule during query and requires computing the 3D surface-based similarity between the query and model molecules. The research presented here specifically focuses on small molecules (tens of atoms, molecular weight around one thousand Daltons), typical of the type used in drug discovery. However, the proposed technique can scale-up equally well to larger molecules like proteins and problems related to their similarity-based structural classification. *The key contributions of this work lie in the similarity formulation being considered and in the efficiency and accuracy of the proposed approach for determining molecular similarity.* The distinctions of the proposed similarity formulation from the more prevalent ones based on searching for sub-structural motifs (sub-structure search) are:

- *Molecular representation*: It is generally accepted that receptors and substrates recognize each other at their molecular surface [8]. Therefore, surface-based descriptors (see Figure 1) provide representational capabilities that are more faithful to the actual nature of molecules than commonly used 2D molecular graphs-based approaches. Such representations have been espoused in the works of many researchers (see for example, [19, 20, 24, 28, 31] and references therein). The proposed similarity formulation allows *direct querying and retrieval* of molecules when they are represented using complex and bio-chemically relevant surface-based descriptors.
- *Query formulation:* In a sub-structure-based search the query is specified as a molecular sub-structure (typically represented as a 2D molecular connectivity graph) and the retrieved molecules are constrained to contain the entire sub-structure specified in the query. This formulation requires the user has to have a clear picture of the structures which are to be retrieved prior to issuing the query [4]. Typically such detailed knowledge is available only when the mechanism of action of the molecule is established in terms of its activity as determined by specific structural fragments. In contrast, “whole molecule” similarity is suitable for exploring structural space [4], generating hypotheses, or querying chemical databases when detailed structure-activity information, at the level necessary for sub-structure querying is unavailable.
- *Applications:* Many biological properties like interaction of molecules with biological membranes or receptor-ligand interactions are mediated by molecular characteristics like geometry, hydrogen bonding, polar molecular surface, electrostatics, and hydrophobicity [20] which are surface-based and/or defined over the entire molecule. Membrane permeation in the human intestine, which is central to absorption of an orally administered drug in humans and blood-brain permeation of molecules are two of the many possible examples of such properties. Similarity formulations, such as those considered in this paper can therefore be used to build models that relate biological properties to the structure of molecules (structure-property models). In contrast, 2D substructure-based similarity is typically ill-suited in the context of studying such biological activity. However, if information linking substructure(s) to specific activity is available then 2D substructure-based similarity can be a more appropriate measure [4].

## 1.1. Problem characteristics and challenges

From a computational and algorithmic perspective, the problem of determining the similarity of molecules when they are represented using complex 3D surface-based descriptors presents some unique challenges which include:

1. *Definition of a standard coordinate system for surface-based molecular representations*: To compare molecules using their surface-based descriptions, it is necessary to have a way of representing the shape of their surfaces. The complexity lies in defining an intrinsic (view independent) coordinate system over the curved molecular surface that maps a point on the curved surface to a point on a standard coordinate system. This is necessary to establish a match between feature distributions corresponding to two molecules (the model and the query). Additionally, such a mapping should be one-to-one between points on the molecular surface and the standard coordinate system, so that it can faithfully represent the relationship between the structure of a molecule and its bio-chemical properties.
2. *Multimodal nature of molecular properties*: Properties like geometry and charge distributions that may be used for describing molecules have entirely different characteristics. For example, while the geometric representation of a molecule is unique, its field-effects are superposition-based.

Thus, structurally different molecules may show similar biological activity (due to similar field-effects). Representation and similarity formulations need to account for such issues.

3. *Extensibility*: Depending on the biological context, different surface-based molecular properties may be involved in determining molecular similarity. This requires a similarity formulation to be extensible beyond the properties it was originally designed for. As a corollary, the formulation should also be able to support assignment of different weightings or importance to the various molecular properties over which similarity is determined.
4. *Molecular pose and conformations:* During a similarity query, the pose of the participating molecules can be arbitrary. Further, each molecule may be represented by one of many energetically minimal configurations, called conformations. On one hand the representation and similarity formulation should be invariant to molecular pose. On the other, it should be sensitive to effects like conformational changes since the bio-chemical behavior of a molecule can vary significantly depending on its conformation.
5. *Query efficiency*: It is typical to conduct molecular similarity queries over large sets ranging from thousands to millions of molecules. The latter order of magnitude is especially common in pharmaceutical and drug discovery settings. It is therefore imperative for similarity determination approaches to be computationally efficient.
6. *Validation*: Although a number of research efforts have focused on the problem of molecular similarity, few (such as [25, 34]) have actually attempted to validate the significance of proposed similarity measures in terms of molecular structure-property relationships [4]. This may be ascribed to various factors that lead to the complexity of such modeling and possible lack of appropriate biological activity data. However, such a validation step is essential in determining how well the underlying bio-chemistry is accounted for in the similarity formulation and its algorithmic solution.

The research presented in this paper approaches the problem of determining molecular similarity in three steps: First, an intrinsic representation for molecules is developed by defining a mapping between the closed molecular surface and a unit sphere. A molecule can then be treated as a collection of distributions defined on the sphere, where each distribution represents a specific molecular characteristic. Salient to this step is the application and extension of results from research in computer vision on the problem of 3D curved-object recognition. The second step is based on the idea that the similarity of two molecules can be obtained by comparing the similarity of the respective molecular property distributions. In doing so, we employ the uniqueness of the spatial distribution of a molecular property on the surface of the encapsulating sphere; this distribution is invariant to the pose (but not the conformation) of the underlying molecule and can be used as a constraint to determine the similarity of two molecules with respect to the given property. The obtained similarity is consequently independent of how the molecules are themselves oriented with respect to each other. To make the similarity score invariant to conformations, a molecule is represented by a set of energetically minimized structures or conformers, over which the similarity is computed. Finally, to address issues related to efficiency, we propose a novel variant of histogram-intersection to compute the similarity of property distributions corresponding to two molecules. The technique of histogram-intersection [36] has been shown to be highly efficient and is used extensively in information and image retrieval. We seek to take advantage of this efficiency when querying large collections of molecules. Our variation on this technique allows determining the similarity of molecular property distributions by considering their spatial characteristics. Essentially, high similarity scores are obtained only if both the statistical and topological natures of the distributions tend to agree.

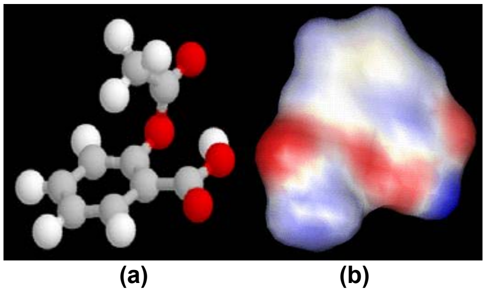

**Figure 1: Representation of the Asprin molecule: (a) Ball-and-stick representation, (b) Surface-based representation**

We begin this paper with an introduction to molecular representations and present an overview of the prior research conducted on the problem of molecular similarity in Section 2. The proposed method is formulated in Section 3. Results from multiple experimental studies are reported in Section 4. The conclusions from this research and directions of possible future work are presented in Section 5.

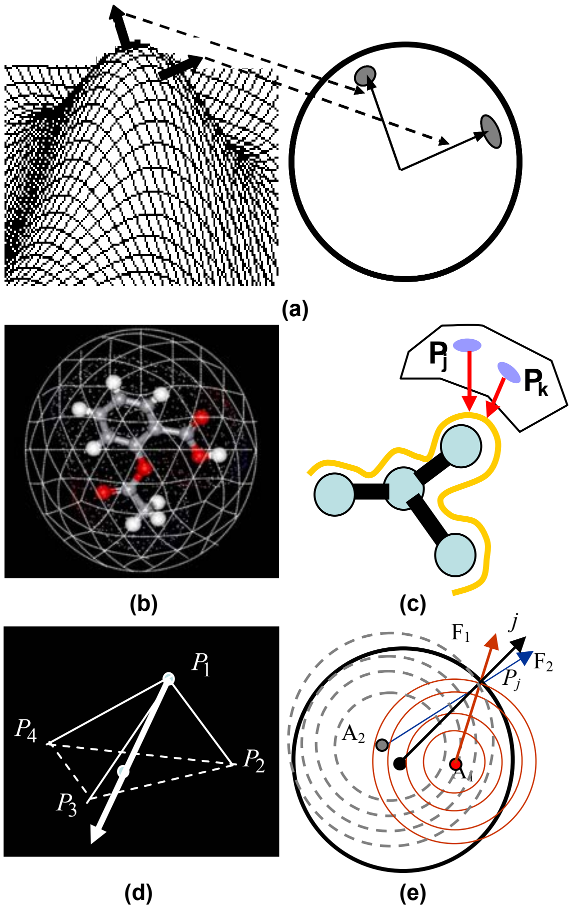

**Figure 2: Illustration of the principle concepts in the proposed molecular representation:** (a) The Gauss mapping, (b) Embedding of a molecule in the tessellated sphere, (c) Representation of the molecular geometry by surface mapping, (d) The local regularity constraint, (e) Superposition-based field measurements

## 2. Molecular representation and prior research

In the simplest form, a molecule may be represented by using its chemical formula. Other representations include the molecular graph, which is based on a connectivity matrix where atoms that participate in chemical bonds are shown to be connected. This graph can also contain information about bond orders and can be used to distinguish isomers (same molecular formula but different topologies). Traversals of such graphs can be used to generate string-based representations for molecules by incorporating atomic symbols and bond-types encountered during the traversal. One such representation is the SMILE string representation of molecules. More complex representations include surface-representations (Figure 1(a)), which are obtained by rolling a probe-atom over the molecule. The molecular surface being defined as the set of points where the surface of the probe atom touches the *van der Walls* surfaces of the atoms constituting the molecule. In spite of its seeming complexity, efficient algorithms, for example [21], requiring $O(nlogn)$ deterministic time and using $O(n)$ space exist for computing surface-based representations. Other representations can be based on the Schrodinger wave equation. However, such quantum representations often require solving non-linear PDEs, which can become prohibitively expensive from any non-trivial molecular system consisting of more than a few atoms.

Research in determining similarity of molecules has been closely tied to molecular representation schemes. Early works in the area like [33] and [39] used variations on the sum of inter-atomic distances. Later approaches have looked at schemes for atom re-labeling to minimize a difference-distance matrix [3], [32] or decomposing the molecular distance and connectivity graphs into subgraphs which are numerically characterized and compared [5]. Another class of methods [19], [24], [28] has defined molecular similarity using surface and field characteristics: First, the field-effects around a molecule are estimated. Then, the orientation of the query or model molecule (a 3D graph), is varied to minimize an RMS error between the field values. Other efforts include the application of *geometric hashing* and its variations [31].

## 3. The proposed method

A pre-requisite for comparing molecules described using surface-based representations is the capability to map points on the curved molecular surface to points on a standard coordinate system. Such a mapping was derived by Gauss [17], by using surface orientations to map points on an arbitrary curved surface to a standard coordinate system defined on a unit sphere. This mapping is formally referred to as the Gauss map and can be defined as follows:

*Definition 1:* Let $G \subset R^3$ be an oriented surface in Euclidean space. Further, let $S$ be a unit sphere, called the Gaussian sphere. The Gauss map $M$ is the mapping $M : G \rightarrow S$, where the surface normal for each point on the surface $G$ is translated to the origin of the sphere $S$ and the end points of each normal lie on the surface of the Gaussian sphere $S$ (see Figure 2(a) for an illustration).

An important derivative of the Gauss map for representing curved surfaces is the *Extended Gaussian Image* (EGI). It is obtained from the Gauss map by assuming that the surface $G$ is evenly sampled into patches and that each surface normal is associated with a single unit of mass which it votes to the corresponding point on the Gaussian sphere. The distribution of the mass on the surface of the Gaussian sphere, obtained in this fashion depends on the shape of the underlying surface and is called its Extended Gaussian Image. Even though the EGI mapping does not incorporate the spatial relationships between the surface patches on $G$, it possesses certain important characteristics that include: (i) If two convex objects have the same EGI, they are provably congruent (the Minkowski theorem [30]), (ii) As an object rotates, its EGI rotates in the same manner, (iii) The EGI mass on the Gaussian sphere is inverse of the Gaussian curvature of the underlying object surface, and (iv) The center of mass of the EGI lies at the origin of the Gaussian sphere.

The properties of the EGI, especially the Minkowski theorem provide the foundations for using it towards representing and comparing surface-based description of objects. However, an inherent problem of EGI-type mappings is their dependence on the Gauss map which is non-unique for non-convex objects. Because of this, more than two points on an object surface may be mapped on the same point on the Gaussian sphere. Unfortunately, many molecules in their stable conformations induce surfaces that are non-convex and therefore the application of techniques from the EGI family is precluded for their representation and matching. In computer vision, researchers have attempted to solve such a problem by using a spherical attribute image (SAI) which is based on iteratively deforming a geodesic surface to fit an underlying object. We build on the above research by replacing the deformable surface mapping by a fast non-iterative mapping-by-probing approach that preserves a local regularity constraint and is sensitive to the underlying surface shape. We begin by placing the molecule inside a semi-regularly tessellated sphere (Figure 2(b)), which for purposes of notational uniformity we will also denote as *S,* but distinguish from the Gaussian sphere based on the context. The placement of the molecule is done such that its center of mass coincides with the center of the sphere. Our approach for tessellating the sphere is based on [23] and involves subdivisions of the triangular sides of a 20-side icosahedron into sub-triangles. At each point of the tessellated sphere we compute three properties to describe the molecule, namely *geometric shape*, *donor field* (due to H-bond donor atoms), and *acceptor field* (due to H-bond acceptor atoms). Our selection of these descriptors is due to their importance in various molecular interactions and the fact that more complex descriptors like polar surface area of the molecule (which influences membrane permeability), are correlated to donor/acceptor fields [37].

In order to motivate the mapping used by us, we begin by first noting that the molecular surface, by its construction is *smooth.* At each point $P_j$ of *S,* the molecular surface is probed to determine the point closest to $P_j$ (see Figure 2(c)) subject to a local regularity constraint illustrated in Figure 2(d)). The regularity constraint requires that the probe vector pass through the centroid of the plane formed by the neighbors of $P_{j.}$ This ensures uniformity in measurement and invariance to translation and rotation of the molecule. The distance to the surface point is then used as an estimate of molecular shape. The measurement of the donor field is done using the following three step procedure:

*Step 1:* The Hydrogen-bond donor atoms in the molecule are identified. Typically these are Nitrogen or Oxygen atoms with hydrogen on them. Other ways of identification like the PATTY-rule [10] can also be used.

*Step 2:* The donor field is defined as an isotropic Gaussian distribution and the field at point $\mathrm{P_j}$ due to an atom at position $\mathrm{X_i}$ having van der Walls radii $r_i$ is defined by Eq. (1) below:

$$f(P_j, X_i) = \left(\frac{a^2}{2\pi r_i^2}\right)^{\frac{3}{2}} \exp\left(\frac{-a^2}{2r_i^2} \mid X_i - P_j \mid^2\right) \qquad (1)$$

In Eq. (1) $a$ is a scale factor for the radii. The value of a=2, for which 90% of the electron density lies inside the van-der walls radius of the atom, is used in all the experiments. The reader may note that a similar approach for field definition has also been suggested in [28].

*Step 3:* At a given tessellate point $\mathrm{P}_{j,}$ having the surface normal *j,* the field strength for each donor atom is computed. The direction of each field is given by a unit vector obtained by joining the corresponding atom to $\mathrm{P}_j$. The resultant field at $\mathrm{P}_j$ is defined as the vector sum shown in Eq. (2) (also see Figure 1(e)):

$$\vec{F}(P_j) = \sum_i [f(P_j, A_i) \times (\overrightarrow{i} \,.\, \overrightarrow{j})] \qquad (2)$$

In this formulation maximum weight is given to those atoms whose field direction coincide with the surface normal at the specific tessellate point, in computing the resultant field. The acceptor field is analogously determined. Typically Nitrogen or Oxygen atoms with a lone pair of electrons are considered as acceptors. At the end of this stage, the molecule is represented by a set of points at each of which three values corresponding to the geometric shape, donor field, and acceptor field are respectively defined.

Similarity of two molecules is defined in terms of the similarity of their property distributions. Histogram intersection provides a rapid way to empirically compute such a similarity. Furthermore, it is highly efficient and is invariant to translation and rotation of the distributions. However, due to the absence of pose information within the molecular descriptors a direct application of histogram intersection is not possible. In Figure 3, we show the intuition behind our solution to this problem. Based on it, our method to obtain the similarity of the distributions can be described as follows:

- For each of the distributions $P_1 \ldots P_K$ used in representing the molecule, define a (fixed) quantization and construct the histograms. Let $H_L$ denote the histogram corresponding to the distribution $P_L, L \in [1, K]$ and let $k$ denote a specific bin in the histogram.
- Cluster the tessellate points having the same value (falling in the same bin $k$) by adjacency of the tessellated surface patches. Compute the centroid for each cluster.
- Compute the distance (constrained to lie on the surface of the sphere) for all pair of centroids.
- Quantize these distances into bins as follows: {[0,1[, [1,2[, . . .[C/2-1, C/2[}, where C is the circumference of the sphere. Compute the histogram for the distance distribution.
- For two distance histograms $D_M$ and $D_I$ (with the bins indexed by $j$) the histogram intersections are computed using Eq. (3). The distance histogram intersection defining the topological similarity of values in bin $k$, denoted by $\gamma_k$ is defined as the average of the two intersection values from Eq. (3) (to ensure symmetry).

$$H(D_I, D_M) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{M_j}} ; H(D_M, D_I) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{I_j}} \tag{3}$$

- The topologically constrained histogram intersection value for a given property distribution $P_L$ between two molecules $M_1$ and $M_2$ is defined by Eq. (4), where the histogram intersection $H(M_1,M_2)\gamma$ is computed, as shown for the general case, in Eq. (5):

$$H_{TC}(M_1, M_2) = \frac{H(M_1, M_2).\gamma + H(M_2, M_1).\gamma}{2} \tag{4}$$

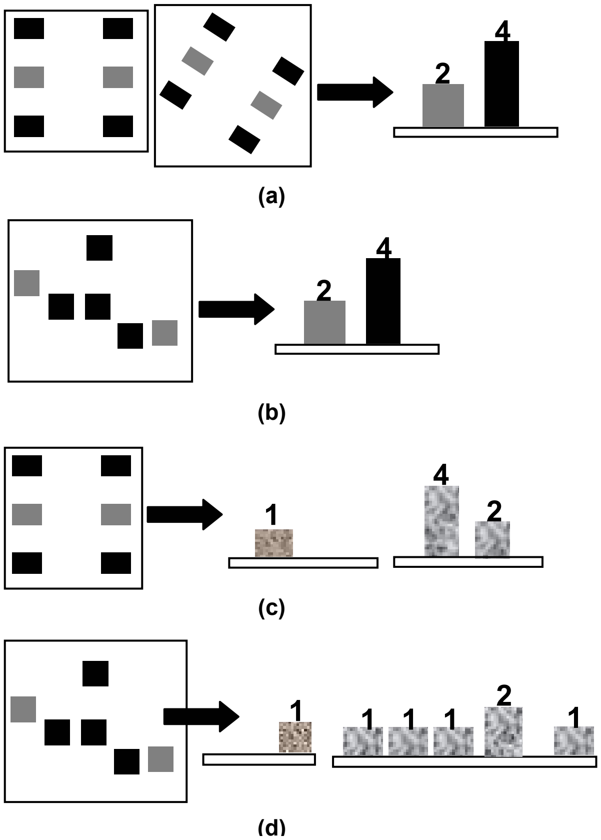

**Figure 3: Intuition behind determining the similarity of molecular property distributions:** Each distribution in (a) consists of four black and two grey elements. The two distributions shown are related to each other by a rotation. The property (color) histogram for these distributions is shown on the right and is invariant to any Euclidean transformations of the distribution. However, the property histogram is, by itself, insufficient to disambiguate distributions as shown by the example in (b). Here the distribution is distinct from those in (a) and yet yields an identical property histogram. The spatial characteristic of the distributions, as typified by the histogram of the pair-wise distances between elements having the same property, is shown in (c) and (d) corresponding to the distributions in (a) and (b) respectively. As can be seen, such a characterization is distinct for each distribution and can be used to determine their similarity. Note that the distance distribution histograms are shown as textured to denote their distinction from the property distribution histograms.

$$H(M_a,M_b)\gamma = \frac{\sum_{j=1}^{K} \min(M_{aj},M_{bj}) \times \gamma_j}{\sum_{j=1}^{K} M_{aj}} \qquad (5)$$

- The full histogram intersection $H_{full}(M_i,M_j)$ between two molecules $M_i$, $M_j$, is the average over all property distributions of the corresponding topologically constrained histogram intersection values. To account for molecular conformations, we define the similarity of two molecules $M_i$ and $M_j$ as the maximum value of the full histogram intersection defined over a set of conformations the molecules can assume. This is formally described in Eq. (6). The conformations for a molecule can be generated by finding the local minima of a non-linear function that defines the energy of a molecular structure. This function contains terms capturing the energy due to inter-atomic interactions as well as due to deviations in bond length, bond angles, and tertiary angles. Typically a standard package like CONCORD [42] is used for this purpose.

$$\begin{gathered} Similarity(M_i,M_j) = \underset{C_i,C_j}{\arg\max}[\mathrm{H}_{\mathrm{full}}(C_i,C_j)] \\ C_i = \{C_i^1, C_i^2, \ldots, C_i^r\}, C_j = \{C_j^1, C_j^2, \ldots, C_j^r\} \end{gathered} \qquad (6)$$

## 4. Experimental evaluations

The efficacy of the approach was tested using three sets of experiments. The experiment design for each set incorporated two stages: The first stage involved a direct application of the method on a data set to solve a specific problem. In the second stage, a comparative study was performed by applying a state-of-the-art research or commercial technique on the same problem and data set. Subsequently the results were analyzed to understand the distinctions and contributions of the proposed approach. The three experiments included: (a) Validation of the method's accuracy in query-retrieval settings, (b) Evaluation of its performance (speed), and (c) Validation through applications in structure-activity modeling problems.

### 4.1. Accuracy in query-retrieval settings

In the first stage, the method was tested in a query-retrieval setting on a subset of 5000 molecules randomly selected from the MDDR collection [41]. The MDDR is a commonly used reference in drug discovery and structural biology and consists of molecules that are either marketed drugs or have reached advanced stages in a drug discovery process. Each of the 5000 molecules was successively used as a query against the rest of the molecules in this set. The query and model molecules were each represented by 20 conformers, i.e. 400 distinct molecular structures were used per similarity computation. Since the proposed method does not require super-positioning of the underlying structures for computing similarity, to distinguish its performance from approaches that do so, a variation of the experiment was performed where the query was represented by 20 novel (distinct from the model) conformers. It may be noted, that for some molecules, 20 novel energetically stable conformers could not be obtained. In such cases, as many novel conformers as could be derived for each specific structure were used. In the second stage of this experiment, for purposes of comparison, the query-retrieval experiments were performed using ISIS [41], a widely used commercial 2D chemical database. ISIS uses structure-keys in conjunction with indexing for answering queries. However, molecular similarity using ISIS is strictly 2D-substructure-based and can not incorporate issues like conformations. The consolidated results from these two stages are presented in Table 1. The first row of the table shows results obtained with ISIS. The second row presents the results obtained with 20 conformers for each of the query and model molecules. The final row shows the accuracy of the retrieval process when distinct conformers (between the query and the model) were employed. Here, the asterisk denotes the aforementioned fact that for some molecules 20 distinct stable conformers were not obtained. In this setting, of the 5000 molecules, 4910 were correctly identified. An analysis of the results obtained in this step indicates that the accuracy of the proposed approach during query-retrieval is comparable to that of ISIS, even though the proposed method addresses the query-retrieval problem in a setting that involves molecular conformations, surface-properties, and superposition-based effects and is much more complex than the 2D structural motif-based search used in ISIS.

| **Method** | **Data Size** | **Number of Conformations** | **Accuracy** |
|---|---|---|---|
| ISIS | 5000 | none | 100% |
| Proposed | 5000 | 20/20 | 100% |
| Proposed | 5000 | 20/20* | 98.2% |

**Table 1: Summary of results from the query-retrieval experiment on the MDDR database**

## 4.2. Evaluation of the performance (speed)

In the second experiment, the computational performance of the proposed approach was tested with respect to the system described in [24] and related works [19, 25]. This selection was based on the fact that both the proposed approach and the one described in [24] seek to define the surface-based similarity between molecules. Their distinctions lie in how the modeling of molecular shape and field-effects are accomplished as well as in how the similarity is computed. Furthermore, our selection was also motivated by the fact that [24] along with its derivatives have been extensively applied in pharmaceutical research settings [19, 24, 25] and the published results as well as our own investigations show it to be amongst the fastest approaches currently available for determining surface-based molecular similarity.

In our experiment, 30 pre-selected, maximally-diverse molecules from the MDDR collection were compared against each other, with 20 conformers for the model and one for the query. Both the systems reported a 100% recognition rate on this subset of molecules. However, the time requirements were significantly different. A graph plotting the time required for the similarity computation with the proposed technique is shown in Figure 4 (left plot with the data points shown as triangles). Figure 4 (right plot) shows a comparison of the performance with the method outlined in [19] (data points obtained with [19] are shown as rectangles). On an average, with the proposed technique 120 conformers were processed (descriptor generation and matching) every second, while with [19], one conformer was processed every two seconds on an SGI Indigo2 machine. Another recent method [28], available commercially, reports speeds of 2 minutes per molecule (on a SUN Ultra-30).

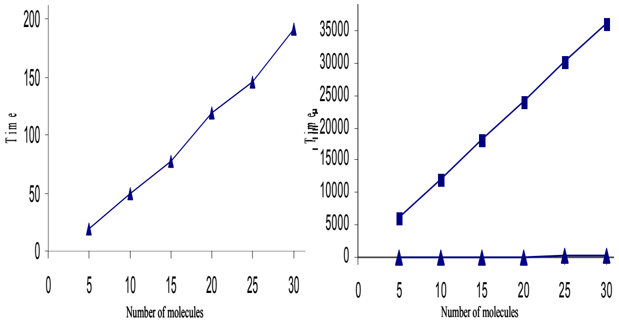

**Figure 4: Computational performance of the proposed method (left) and comparison with [19] (right)**

## 4.3. Validation through application in structure-activity models

Similarity information from the proposed technique was applied for building a structure-activity model for modeling and predicting human intestinal drug absorption. As mentioned in the introduction, intestinal absorption is critical for the success of orally administered drugs. Furthermore, it constitutes one of the primary reasons for expensive late-stage failures in drug discovery. Therefore, robust models linking molecular structure to intestinal absorption can have significant applications in pharmaceutical research and development. The data set used in this experiment consisted of 30 compounds that were tested for human intestinal permeation using the Caco-2 assay. The Caco-2 (human colon adenocarcinoma cell line) provides a close approximation of *in vivo* absorption and can be used to model the epithelial cell layer barrier and absorption from the intestinal lumen to the blood stream. The assay protocol used in this experiment was designed to measure uni-directional flux and all compounds were analyzed at identical initial concentrations. The range of measured values was between 0.0% (no permeation) to 2.8% (maximum permeation) flux units.

The descriptor design for the structure-activity model involved computing the similarity of the participating molecules with a predefined set of thirty molecules, called the *characteristic molecules*. The computed octanol-water partition coefficient (clogP) was used as an additional descriptor. Together, this yielded a 31-dimensional descriptor space. The central idea, behind the notion of characteristic molecules is closely related to the concepts of vector quantization [12] and involves tessellation and quantization of the *d*-dimensional molecular descriptor space *D* into a finite subset *C* of the *d*-dimensional space. Formally, this process can be denoted as a mapping *Q*, which is defined as:

$$Q : D^{d} \rightarrow C, C = \{c_{1}, c_{2}, ..., c_{m}\} \wedge \forall j, c_{j} \in D^{d} \qquad (7)$$

Representing the molecules in terms of their similarity to the characteristic molecules introduces an implicit dimensionality reduction. We refer the reader to [35] for details on such dimensionality reduction approaches in structure-activity modeling. Two measures were used for evaluation of the results. The first is a ratio-scale measure called cross-validated $r^2$ and shows how well the model predicts data that was *not used* during model construction. The definition of this measure is presented as Eq. (8):

$$r^2 = 1 - \frac{\sum_i (V_i - P_i)^2}{\sum_i (V_i - \bar{V})^2} \qquad (8)$$

In Eq.(8), $V_i$ is the experimentally determined property of the molecule $i$, $P_i$ is its predicted property, and $\bar{V}$ is the mean experimental property value. The second measure is an ordinal measure called Kendall's $\tau$, that shows how well the *ordering* of the data is preserved during prediction by the model. A perfect value of $\tau = 1$ is obtained when the predicted order coincides with the order as determined by actual experimental property values. This measure is computed, for $n$ molecules as depicted in Eq (9).

$$\tau = \frac{correct\_ordering - incorrect\_ordering}{n(n-1)/2} \qquad (9)$$

The ordinal measure is important because the ordering (or prioritization) of the molecules is typically more robust to experimental variability than pure error-based measures. Using a combination of the above measures, therefore allows evaluation of a model both in terms of its numeric predictive accuracy, and in terms of how well it can maintain prioritization of the molecules.

Model construction was done using the 20 training molecules. As part of the descriptor selection step, the complete cross-correlation matrix of the descriptors was computed and the top eight least correlated descriptors selected. A backpropogation network with a single hidden layer was used to learn the (empirical) mapping between the molecules as defined by the 8-dimensional feature vector and their permeability values. Learning was stopped when the cross-validated error became lower than a predefined threshold. Figure 5(a) shows the leave-one-out cross-validated performance of the learning model when the similarity relative to the characteristic molecules was computed using the algorithm in [19]. In this setting, one compound was randomly excluded from the training set and the remaining compounds used to learn a model that predicted the permeability for the excluded compound. For the model that was learnt, the value for cross-validated $r^2$ equaled 0.64 and Kendall's $\tau$ equaled 0.29. Figure 5(b) shows the performance of the learned model in a leave-one-out cross-validation setting for the training set when the proposed similarity measure was used. In this case the cross-validated $r^2$ equaled 0.97 and the value for Kendall's $\tau$ was 0.65. It should be emphasized that in both experiments an identical learning algorithms (single hidden layer neural network with back-propagation) was used and the only distinction laid in the similarity values (due to the corresponding approaches for determining them). We also note that in both the experiments, the values for Kendall's $\tau$ typically tended to be low. In case of the proposed approach this was primarily because the original data had compounds that showed 0%

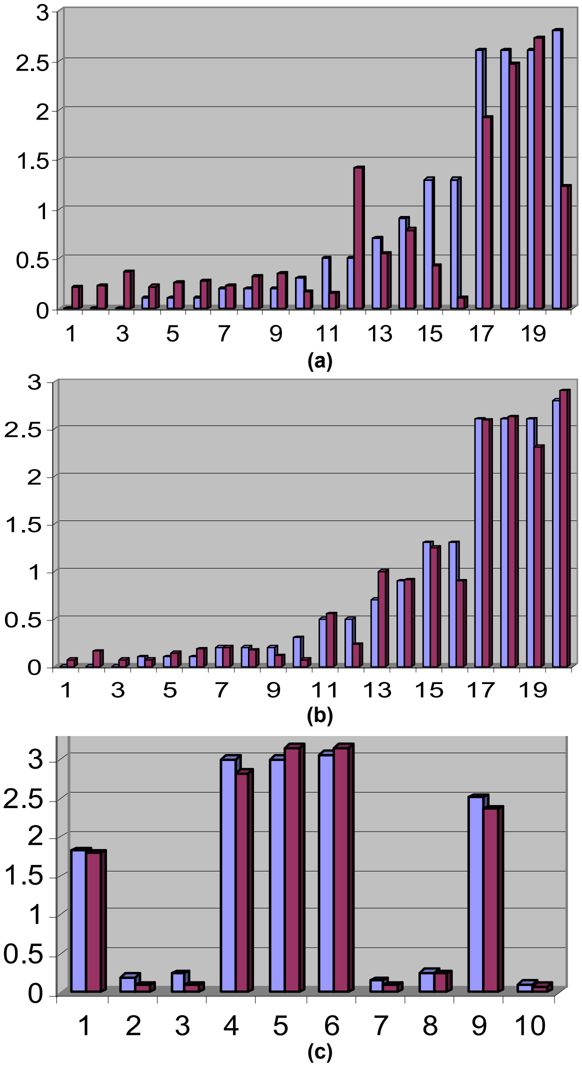

**Figure 5: Performance in structure-activity modeling.** Permeation of each compound is presented by two adjacent bars with predicted values shown by lighter shaded bars on the left and measured values by darker shaded bars on the right: (a) Leave-one-out prediction results from the learning model obtained by using similarities determined by the method [19]. (b) Leave-one-out results using similarity determined by the proposed algorithm. (c) Performance on the test set

absorption (no absorption). However, the model assigned very low (albeit non-zero) absorption values to these molecules, thereby leading to a change of ordering amongst them. However, the model based on similarities derived using [19], showed ordering errors across the entire spectrum of values. Finally, (c) shows the performance of the structure-activity model on the test set of 10 molecules. The prediction results for the test set is also tabulated and presented in Table 2.

| Compound ID | Predicted Permeability | Actual Permeability |
|---|---|---|
| 000753 | 1.79 | 1.83 |
| 091217 | 0.09 | 0.20 |
| 322835 | 0.09 | 0.24 |
| 422025 | 2.83 | 3.01 |
| 489595 | 3.16 | 3.01 |
| 525792 | 3.15 | 3.06 |
| 531746 | 0.09 | 0.15 |
| 598738 | 0.24 | 0.26 |
| 696705 | 2.36 | 2.51 |
| 835218 | 0.08 | 0.11 |

**Table 2: Predicted and measured permeability values for the molecules in the test set. All permeability values are in %flux units. The compound ID is a unique numeric identifier for each molecule and has no relation to the molecular structure**

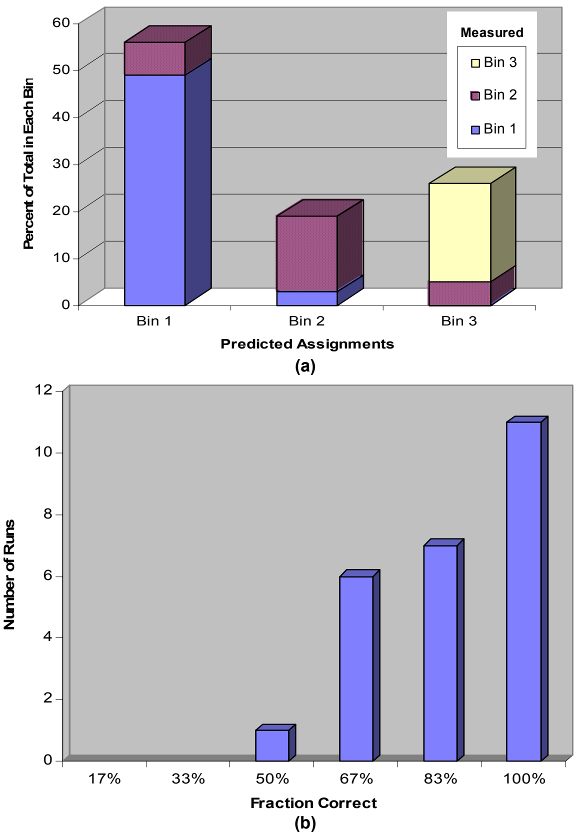

**Figure 6: Analysis demonstrating the robustness of the structure-activity model using leave-n-out cross-validation settings:** (a) **overview of the correctness of bin assignments,** (b) **distribution of the prediction results**

In Figure 6, we present analysis of the method's performance in leave-n-out cross-validated experiments. In the leave-n-out setting, 7 of the 20 training molecules were randomly excluded; the remaining 13 molecules were then used to build a model that predicted the values for the 7 excluded molecules. Since a significant number of samples can get left out in such a formulation, it can provide a good indication of the robustness of the model. To simplify the presentation, the absorption values and predictions are grouped into three bins: bin 1 corresponded to molecules exhibiting poor absorption (defined to be less than 0.5% flux units), bin 2 corresponded to medium absorption (between 0.5% and 1.0% flux), and bin 3 corresponded to molecules that showed high permeation (greater than 1.0% flux). The results shown in Figure 6 are based on the performance of the model in 25 iterations of the leave-n-out experiment. The bar-chart in Figure 6(a) shows the number of incorrect bin assignments that were made: Over the 25 iterations, 85% of the overall bin assignments were correct and in 15% of the assignments, an error of one adjacent bin was observed (i.e. a compound with low absorption got assigned to the medium absorption bin or vice-versa, or a medium absorption compound was assigned to the high absorption bin). However, in none of the iterations, was a low absorption compound predicted as a highly absorbed one or a highly absorbed compound predicted to be a poorly absorbed one. Figure 6(b), presents the distribution of the prediction results across the 25 iterations of the leave-n-out cross-validation experiment: 11 of the 25 iterations resulted in perfect bin assignments and 7 of the 25 iterations had 83% correct bin assignments. Further, 6 of the iterations had 67% accurate assignments and only one of the 25 iterations had 50% accuracy in bin assignments. These statistics indicate the high consistency in the prediction performance of the model across variations in the training set. The reader may also note that all non-accurate assignments involved miss-classifications by no more than one adjacent bin as described earlier.

## 5. Conclusions

In this paper, we considered the problem of defining similarity between molecules based on complex surface-based representations. Such representations capture the physics of the molecules better than commonly used molecular-graph-based approaches and can therefore have significant relevance in molecular query-retrieval, similarity-based exploration of structural space, and in construction of robust structure activity models. We have proposed a novel approach for defining a standard coordinate system for describing complex surface-based molecular descriptions. In this approach the multimodal nature of molecular properties can be accounted for. For computing the similarity of molecules defined in this coordinate system, we propose a novel technique to compare the molecular property distributions using a topologically-constrained formulation of histogram intersection. In this formulation, the numeric as well as the spatial (topological) characteristics of molecular surface-based property distributions can be accounted for. Experimental results indicate that the similarity formulation can be used for highly-accurate query-retrieval and outperforms, in terms of computational speed, existing research and commercially available solutions for determining surface-based molecular similarity. The proposed approach was also validated by applying it in building structure-activity models for complex bio-chemical properties.

The explosive increase in the size of structural (both protein and small molecule) databases and their importance in contemporary biology and drug discovery, underlines the necessity for fast and accurate techniques for determining molecular similarity. The method presented in this paper is a fundamental advancement in this area and can be expected to find broad applicability in applications across biological and pharmaceutical research where molecular similarity plays an important role.

## 6. References

[1] Ajay, P. Walters, and M. Murcko, "Can We Learn To Distinguish Between 'Drug-like' and 'Nondrug-like' Molecules", Journal of Medicinal Chemistry, Vol. 41, pp. 3314-3324, 1998

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, *Book Title*, Publisher, Location, Date.

[2] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson, "A Computer Vision Based Technique For Sequence Independent Structural Comparison of Proteins", Protein Engineering, Vol. 6, pp. 279-288, 1993.

[3] Barkat T. and Dean M., *J. of. Comp.-Aided Mol. Design, 4, 107, 1991*

[4] J. Barnard, G. Downs, and P. Willett, "Descriptor-Based Similarity Measures for Screening Chemical Databases", pp. 59-80 in Virtual Screening for Bio-Active Molecules, Methods and Principles in Medicinal Chemistry, Vol. 10, Eds H-J Bohm and G. Schneider, Wiley-VCH, 2000

[5] G. Bemis, and I. Kuntz, "A fast and efficient method for 2D and 3D molecular shape description", *J. of Comp. –Aided Mol. Design, 6, 607-628, 1992*

[6] M. Bogyo et al. "Selective Targeting of Lysosomal Cysteine Proteases with Radiolabled Electrophilic Substrate Analogs", Chem. Biol. Vol. 7, No. 1, pp 27-38

[7] "Virtual Screening for Bio-Active Molecules: Methods and Principles in Medicinal Chemistry", Eds. H-J Bohm and G. Schneider, Wiley-VCH, 2000

[8] G. Bravi, E. Gancia, D. Green, and M. Hann, "Modeling Structure-Activity Relationships", in *in Virtual Screening for Bio-Active Molecules, Methods and Principles in Medicinal Chemistry, Vol. 10, Eds H-J Bohm and G. Schneider, Wiley-VCH, 2000*

[9] H. Briem and I. Kuntz, "Molecular Similarity Based on Dock-Generated Fingerprints", Journal of Medicinal Chemistry, Vol. 39, pp. 3401-3408, 1996

[10] B. Bush and R. Sheridan, "PATTY: A programmable Atom Typer and Languagefor Automatic Classification of Atoms in a Molecular Database", *J. Chem. Inf. Comp Sci., 33, 756-762, 1993*

[11] H. C. Kolb, "Click Chemistry: Diverse Chemical Function from a Few Good Reactions", Angew. Chem. Int. Ed. Vol. 40, No. 11, pp. 2004-2021, 2001

[12] V. Cherkassky and F. Mulier, "Learning From Data", Wiley Inter-Science, 1998

[13] R. Cramer, et. al., "Prospoective Identification of Biologically Active Structures by Topomer Shape Similarity Searching", *J. Med. Chem,, 42, pp. 3919-3933, 1999*

[14] B. Cravatt and E. Sorensen, "Chemical Strategies for the Global Analysis of Protein Function", Curr. Opin Chem Biol, Vol. 4, No. 6, pp. 663-668

[15] P. Crivori, G. Cruciani, P-A Carrupt, and B. Testa, "Predicting Blood-Brain Barrier Permeation From Three-Dimensional Molecular Structure", Journal of Medicinal Chemistry, Vol. 43, pp. 2204-2216, 2000

[16] M. Deshpande, M. Kuramochi, and G. Karypis, "Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds", ICDM 2003

[17] K. F. Gauss, "General Investigation of Curved Surfaces", Raven Press, New York, 1965

[18] J. F. Gibrat, T. Madej, S. H. Bryant, "Surprising Similarities in Structure Comparison", Current Opinions in Structural Biology, Vol. 6, pp. 377-385

[19] A. Ghuloum, C. Sage, A. Jain, "Molecular Hashkeys: A Novel Method for Molecular Characterization and its Application for Predicting Important Pharmaceutical Properties of Molecules", *J. Med. Chem, 42, 10, pp 1739-1748, 1999*

[20] W. Guba and G. Cruciani, "Molecular Field-Derived Descriptors For The Multivariate Modeling of Pharmacokinetic Data", in Molecular Modeling and Prediction of Bioactivity, K. Gundertofte and F. Jorgensen Eds, Kluwer Academic/Plenum Publishers, New York, pp. 89-94, 2000

[21] D. Halperin and M. Overmars, "Spheres, Molecules, and Hidden Surface Removal", Proc. 10th Annual Symposium on Computational Geometry, pp. 113-122, 1994

[22] D. K. Han et al., "Quantitative Profiling of Differentiation Induced Microsomal Proteins Using Isotope-Coded Affinity Tags and Mass Spectrometry", Nature Biotechnology, Vol. 19, No. 10, pp. 946-951, 2001

[23] M. Hebert, K. Ikeuchi, H. Delingette, "A Spherical Representation for Recognition of Free-Form Surfaces", IEEE Trans. On PAMI, 17, 7, pp 681-689, 1995

[24] A. Jain, K. Koile, and D. Chapman, "Compass: Predicting Biological Activity from Molecular Surface Properties. Performance Comparison on a Steroid Benchmark", J. Med. Chem., 37, pp 2315-2327, 1994

[25] A. Jain, T. Dietterich, R. Lathrop, D. Chapman, R. Critchlow, B. Bauer, T. Webster, and T. Lozano-Perez, "Compass: A Shape-Based Machine Learning Tool for Drug Design", Journal of Computer Aided Molecular Design, Vol. 8, pp. 635-652, 1994

[26] E. Katchalski-katzir, I. Shariv, M. Eisenstein, A. Friesem, C. Aflalo, and I. Vakser, "Molecular Surface Recognition: Determination of Geometric Fit Between Proteins and Their Ligands By Correlation Techniques", Proc. National Academy of Sciences USA, Vol. 89, pp. 2195-2199, March 1992

[27] G. Klebe, U. Abraham, and T. Mietzner, "Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity", Journal of Medicinal Chemistry, Vol. 37, pp. 4130-4146, 1994

[28] P. Labute and C. Williams, "Flexible Alignment of Small Molecules", *J. Med. Chem., 44, 10, pp. 1483-1490, 2001*

[29] I. Lukovits, *J. of Chem. Inf. And Comp. Sci., 31, pp 503, 1991*

[30] L. A. Lysternik, "Convex Figures and Polyhedra", Dover Publications, New York, 1963

[31] R. Norel, D. Fischer, H. Wolfson, and R. Nussinov, "Molecular Surface-Recognition by a Computer Vision-Based Technique", *Protein Engineering, 7, 1, pp 39-46, 1994*

[32] M. Papadopoulos, P. Dean, *J. of. Comp.-Aided Mol. Design, 5, 119, 1991*

[33] Randic M., *J. Am. Chem. Soc., 97, 6609, 1975*

[34] R. Sheridan and D. Miller, "MCS-Based Similarity", Journal of Chemical Information and Computer Science, Vol. 38, pp. 915-924, 1998

[35] R. Singh, "Issues in Computational Modeling of Molecular Structure-Property Relationships in Real-World Settings", Proc. Conference on Systemics, Cybernetics, and Informatics, 2004

[36] M. Swain and D. Ballard, "Color Indexing", *Int. J. of Comp. Vision, 7, 1, pp 11-32, 1991*

[37] D. Veber, S. Johnson, H-Y. Cheng, B. Smith, K. Ward, and K. Kopple, "Molecular Properties that Influence the Oral Bioavailability of Drug Candidates", J. Med. Chem, 45, pp 2615-2623, 2002

[38] P. Walters, M. Stahl, and M. Murcko, "Virtual Screening – An Overview", Drug Discovery Today, Vol. 3, No. 4, pp. 160-178, 1998

[39] H. Wiener., *J. of Am. Chem. Soc., 69, pp 17-20, 1947*

[40] www.daylight.com

[41] http://www.mdli.com

[42] www.tripos.com