JOURNAL OF CHEMICAL INFORMATION AND MODELING

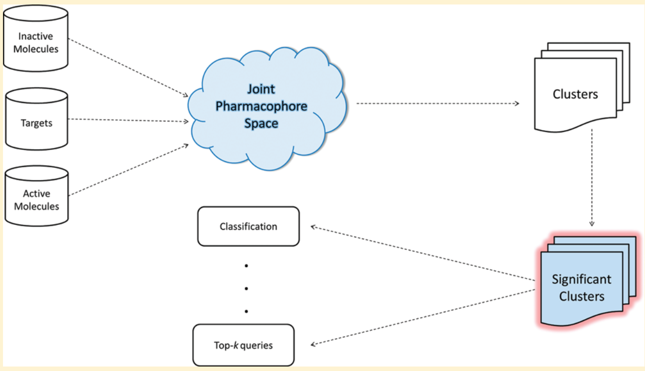
ARTICLE

pubs.acs.org/jcim

Novel Method for Pharmacophore Analysis by Examining the Joint Pharmacophore Space

Sayan Ranu* and Ambuj K. Singh*

Department of Computer Science, University of California, Santa Barbara, Santa Barbara, California, United States



ABSTRACT: We propose a novel method for pharmacophore analysis by examining the *Joint Pharmacophore Space* of chemical compounds, targets, and chemical/biological properties. The proposed approach is a notable deviation from existing techniques that analyze compounds on a target-by-target basis, aimed at extracting and optimizing a specific pharmacophore. The underlying geometry of the pharmacophores is responsible for binding between compounds and targets as well as properties of compounds such as Blood Brain Barrier permeability. The identification of this joint space enables us to cluster and classify similar pharmacophores based on geometric arrangements, analyze the diversity of this space, ascribe positive/negative properties to the subspaces, and query and mine a database of compounds for presence or absence of activity. Extensive experiments are carried out to validate the presence of subspaces that uniquely identify geometric configurations conforming to certain biological activities. The discriminative potential of these subspaces is also verified by employing them as a molecular descriptor. Empirical results show promising performance in terms of classification quality highlighting the utility of mining the joint pharmacophore space.

INTRODUCTION

Despite steady and significant increases in R&D spending, the number of new drug applications and approvals has been, at best, flat. The low productivity of current target-driven approaches to drug discovery has been ascribed to a number of reasons including limited focus to a single target, and undesirable effects such as toxicity and low efficacy that are discovered too late in the discovery process.[1] As a result, current interest is shifting toward evaluating biological properties at the onset and attempting to gain a global understanding of the binding activity between compounds and targets.[2,3]

There have been a number of attempts to understand the relationship between drug chemical structures and target proteins. In one such study, Yamanishi et al.[4] develop a supervised method to infer unknown drug-target interactions by integrating chemical space and genomic space. The authors make predictions for four classes of important drug-target interactions involving enzymes, ion channels, GPCRs, and nuclear receptors. The method measures chemical similarity in the graph domain by considering the size of the largest common subgraph between two compounds. Keiser et al.[5] compare protein families based on the chemical structure (Tanimoto coefficient) of the sets of ligands that bind to them. Yildirim et al.[6] synthesize a global drug-target network consisting of different protein classes with a bipartite graph representation, but the authors do not use the chemical structure information in this analysis.

A number of computational approaches have also been developed to analyze and predict compound-protein interactions. A commonly used method is docking.[7,8] However, docking requires 3D structures of proteins and so cannot be used on a large scale. Wale and Karypis[9] develop a technique for "target fishing" (finding all possible targets for a given compound) by analyzing the target-ligand activity matrix using Support Vector Machines (SVM) and perceptrons. Here, each chemical compound is represented by a frequency vector of topological descriptors. Other techniques for such prediction have used nearest-neighbors,[10] Bayesian models,[11] and neural networks.[12]

Closer to drug discovery, structure−activity relationships (SAR) have been used to guide the iterative optimization of drug leads. Recently, scientists have focused on improving SAR models by considering additional information besides the known ligands to the target under consideration. These approaches include an iterative SVM where training examples at the decision boundary are added to the training set[13] and techniques that refine the SAR score using neighboring protein−ligand pairs in the joint space.[14,15] The latter group of chemogenomics techniques differ from each other based on the descriptors they use for representing the target, ligand, or the complex or the machine learning method used for prediction.[14−24] This thread of research again considers global information.

Pharmacophore based screening has also witnessed significant activity in computer aided drug design. A pharmacophore is a spatial arrangement of chemical features that defines a pattern

Received: December 21, 2010
Published: April 13, 2011

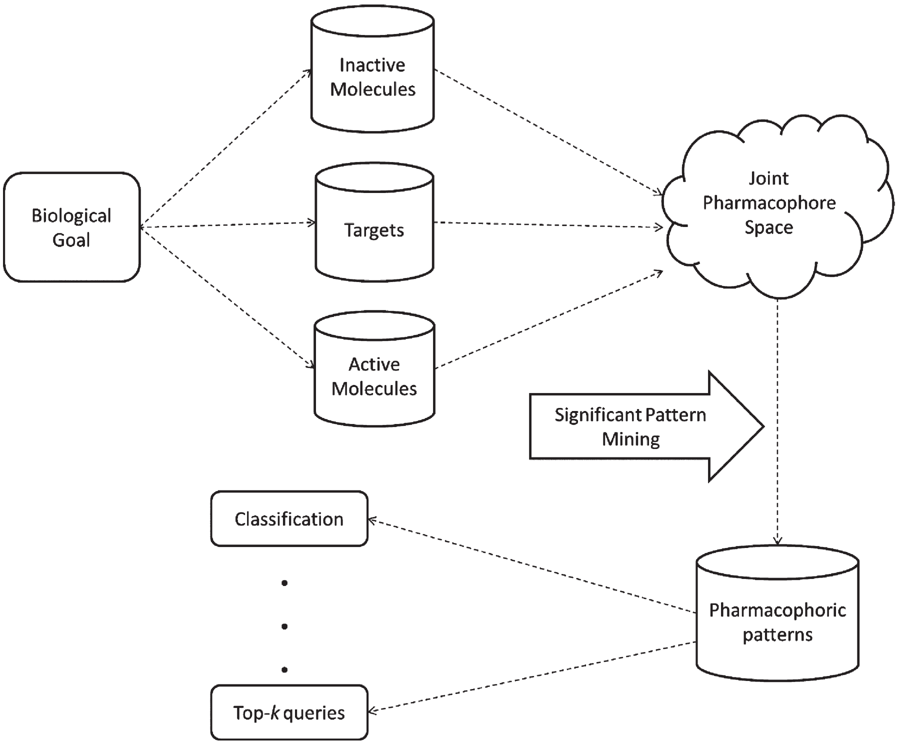


Figure 1. Outline of the proposed approach.

essential for biological activity. Chemical features taken into account in defining pharmacophores usually include hydrogen bond donor/acceptor, charge, hydrophobicity, and aromacity. The geometry of the arrangements of pharmacophores is responsible for binding between compounds and targets as well as properties of compounds such as Blood Brain Barrier (BBB) permeability[25] and toxicity. A number of excellent tools including Phase,[26,27] Catalyst,[28,29] LigandScout,[30] and MOE[31] are available for discovering pharmacophores based on a set of actives (and inactives) against a target (usually with an unknown structure) and searching a database for compounds matching the pharmacophore.

A key weakness of existing pharmacophore based techniques is however its ability to analyze compounds only on a target-by-target basis, aimed at extracting and optimizing a specific pharmacophore. Such an approach is limited in terms of the search space it can investigate in the drug discovery process. Often, multiple pharmacophoric targets need to be analyzed in search for drugs against diseases such as cancer or AIDS. In this paper, we attempt to address this weakness.

We design a new technique that eliminates the need to optimize pharmacophores against a specific target. Figure 1 outlines the proposed approach. We define a *Joint Pharmacophore Space (JPS)* of chemical compounds, targets, and physicochemical/biological properties using the 3D geometry of pharmacophoric features and mine this space directly to identify pharmacophoric patterns. The identification of similar pharmacophores based on geometric arrangements allows us to ascribe positive/negative properties (such as BBB permeability or hERG receptor activity[32]) to different subspaces and define structure-based filters early in the drug discovery process. The proposed work is unique in that we examine the joint space of pharmacophores by considering the conformations of all known actives against any target. Such an approach allows us to take results from a cell-based assay and deconvolute them into separate activity subspaces, each of which could potentially be responsible for a separate binding. These active subspaces can then be queried to find independent groups of active compounds that can be optimized independently. Such a joint space promises to be a preferred beginning investigation point for medicinal chemists.

To summarize, the paper makes the following contributions to the field of computer aided drug design:

- We formulate the unique definition of a *Joint Pharmacophore Space (JPS)* by utilizing the 3D geometry of pharmacophoric features for all actives against multiple targets.
- We demonstrate novel data mining applications in this joint space by successfully identifying subspaces that show statistically significant binding activity. This is performed by clustering pharmacophoric features of compounds in the geometric space and identifying clusters that correlate with a certain biological activity.
- The application of statistically significant clusters is demonstrated by using representative pharmacophoric features as geometric keys to convert molecules into feature vectors. As shown later, the proposed descriptor based on significant clusters outperforms Daylight fingerprints[33] and 3-point pharmacophore fingerprints[34] in molecular classification.

METHODS

Extraction of Pharmacophoric Features. The concept of pharmacophores is based on the kinds of interactions that take place between a set of small molecule ligands and a protein receptor. Typically, low-energy conformations of a molecule are generated, and different pharmacophoric features of interest such

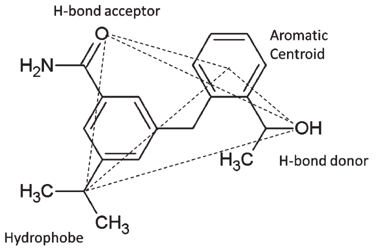


Figure 2. Triplets extracted from the shown pharmacophoric features.

as hydrogen bond donors and acceptors, aromatic rings, hydrophobic cores, and groups with positive and negative charges are extracted. While each of these features plays a role in the binding activity, the exact requirement for a binding to occur typically depends on the presence of multiple such features and the interfeature geometric distances. At the same time, it is more likely that only a part of the molecule takes active participation in the binding activity rather than the entire structure. As a result, two structurally dissimilar molecules might have affinity toward a similar binding activity, if they share the local structure that is required for the binding. Thus, to model this phenomenon, we need to extract features that are local in nature but at the same time are able to capture the interfeature dependencies.

Figure 2 demonstrates our approach of extracting local geometric features. First, we identify all pharmacophoric features in the conformation of a molecule and then extract all possible triplets to characterize the conformation in the geometric space. Figure 2 contains four such triplets. In the geometric space, these pharmacophoric triplets take the shape of triangles and can be thought of as the basic building blocks of any local structure that is required for a pharmacophore model. Specifically, even if the local structure for a binding consists of more than three pharmacophoric features, it can be reconstructed by joining the triplets. The advantages of working with triplets are computational efficiency of the ensuing analysis, and minimality, i.e., three pharmacophoric points are usually the minimum number used in pharmacophores. Similar approaches of working with pharmacophoric triplets have been studied before.[34−36] Such triplets have been used to generate "three-point" pharmacophore fingerprints for molecular analysis. We compare our technique to such fingerprints later in the section discussing the results.

As can be seen in Figure 2, four triangles can be extracted from the molecule. Each triangle is associated with two pieces of information: the 3D coordinates of its vertices in the geometric space and the *triangle type*.

Definition 1 TRIANGLE TYPE: *The type of a triangle is formed by concatenating the types of the pharmacophoric features of its three vertices in ascending order. An example is shown in Figure* 3.

Once the triangles are extracted, they are further grouped into sets based on their types such that all triangles in a group are *mappable* to each other.

Definition 2 TRIANGLE MAPPABILITY: *Two triangles are mappable to each other if and only if a one-to-one mapping can be established between the vertices of the triangles. Due to the unique definition of the triangle type, two triangles are mappable to each other only if they are of the same type.*

An illustration of this grouping is shown in Figure 3 where the molecule has four pharmacophoric features. Like in Figure 2, four triangles can be extracted from the molecule. The types of each of these triangles are shown in Figure 3. The triangles are then grouped into three sets where the group <acceptor, cation, donor> contains two triangles while the other two groups contain a triangle each. The rationale behind grouping triangles into sets is to keep track of triangles that are comparable to each other. More specifically, a similarity or distance between two triangles can be computed only if they are mappable to each other.

The above feature extraction scheme provides a platform to characterize molecules using local pharmacophoric features in the geometric space. Next, we develop methods to analyze the joint pharmacophore space of molecules to identify interesting subspaces that correlate with a certain binding activity. We assume we have a data set of molecules that have been assayed against multiple targets but toward a common biological activity. Further, each molecule in the data set is tagged as either *active* or *inactive*.

Mining Statistically Significant Subspaces. Figure 4 outlines the workflow of mining significant subspaces from the joint pharmacophore space. Given a data set of molecules along with their class labels, we first extract the pharmacophoric triangles from conformations of each molecule. As a result, we transform a database of molecules to a database of triangles. Next, all triangles in the database are grouped into sets based on their types. Once grouped into sets, we cluster triangles in each group and analyze them to identify the clusters that are statistically significant. The goal of the clustering process is to mine geometric structures in the joint pharmacophore space and check whether a subspace is discriminative toward a specific binding or biological activity. At a high level, we annotate the different subspaces with specific chemical/biological properties.

Clustering in the Joint Pharmacophore Space. There are two main components in the clustering process. First, we need a measure to accurately quantify the similarity or distance between two mappable triangles. Second, assuming we have such a distance measure, we need a clustering algorithm that can operate on distance matrices. We need to go to the domain of distance matrices since it is not possible to map triangles to points in the vector space. Although there are hierarchical and graph-based clustering algorithms[37] that can cluster objects based on distance matrices, for our purposes, it is more desirable to use a density-based clustering algorithm. Thus, we focus on devising techniques for the above two components of the clustering procedure.

To accurately quantify the distance between two triangles, we use the Kabsch algorithm.[38] The Kabsch algorithm is a method for calculating the optimal rotation matrix that minimizes the root mean squared deviation (rmsd) between two sets of points. With this algorithm, we compute a distance matrix for each triangle group. Next, to cluster objects using the distance matrix, we use k-medoid clustering. The clustering algorithm proceeds in a manner similar to k-medoid clustering in vector spaces, except for the process of computing the cluster center. In a vector space, the cluster center is computed by simply taking the average of the coordinates of each object in a cluster. However, that is not possible in our case since each object is a triangle.

Algorithm 1 (see Chart 1) presents the pseudocode of the algorithm. Given the desired number of clusters, k, k randomly chosen triangles are assigned as cluster centers (line 1). Next, all remaining triangles are assigned to the cluster with the closest center (lines 3−5). Now, the cluster center for each cluster is recomputed. This is performed by choosing the triangle with the minimum average distance to all other triangles in the same

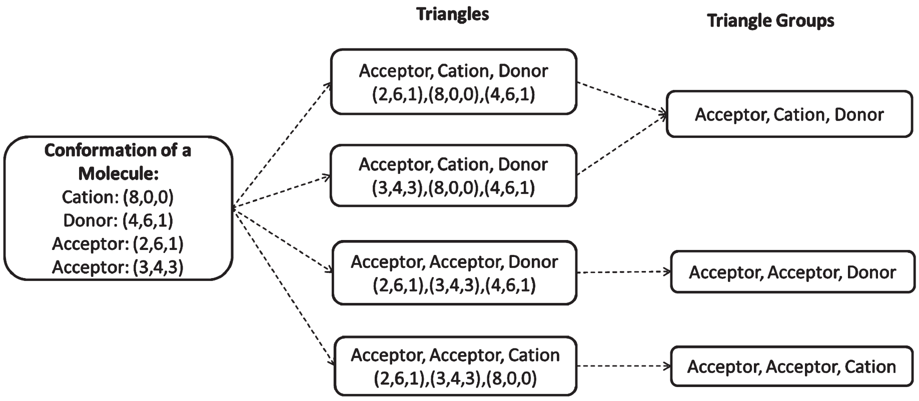


Figure 3. An illustration of how local pharmacophoric features are extracted from a molecule. The molecule has four pharmacophoric features that are defined using the (x,y,z) values. Four different triangles can be formed using these features. Further, the triangles can be grouped into three classes.

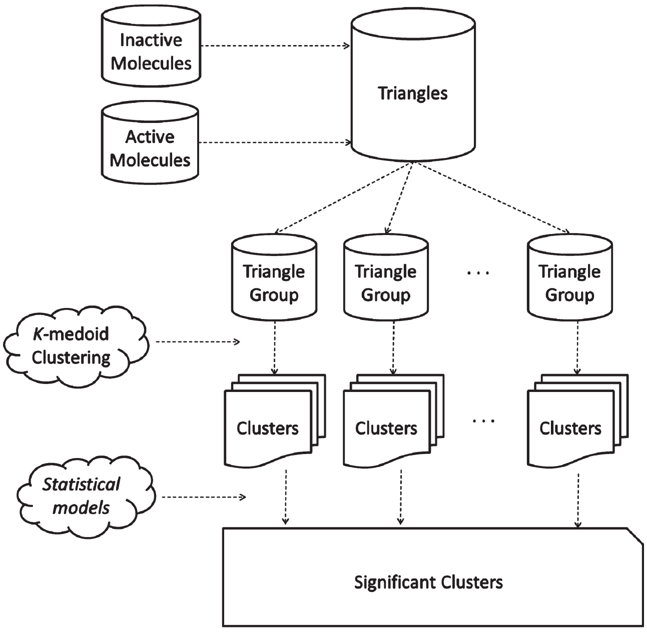


Figure 4. Outline of the proposed approach to mine significant subspaces from the joint pharmacophore space.

cluster (lines 6−7). Statistically, the new cluster can be thought of as the median of the cluster. Once the new cluster centers are computed, the remaining triangles are reassigned to appropriate clusters. The process continues in an iterative manner until the cluster centers converge (line 8). Finally, the clustered triangles are returned (line 9).

Identifying Significant Clusters. Once the joint pharmacophore space is clustered, we analyze the clustered subspaces and evaluate them for biologically/chemically useful properties. More specifically, if the distribution of triangles from conformations of active molecules in a cluster deviates significantly from the expected ratio, then the cluster is discriminative in nature. Thus, to identify such clusters, we develop methods to analyze the statistical significance of a cluster. Statistically significant clusters can then be applied for higher level mining tasks such as molecular classification and top-k similarity queries. Please note that the variable k referred in top-k is independent of the variable k in k-medoid clustering.

First, we formalize the idea of *positive* and *negative* clusters. Given a data set $\mathbb{D}$ of molecules, let $\mathbb{A}$ be the set of active molecules in $\mathbb{D}$. Further, given a cluster of triangles $\mathbb{C}$ originating from conformations of molecules in $\mathbb{D}$, let $\mathbb{P}$ be the set of triangles from conformations of active molecules in $\mathbb{C}$.

Definition 3 Expected Ratio: *The expected ratio, r, is defined as the expected ratio of triangles originating from conformations of active molecules in any given cluster $\mathbb{C}$. Mathematically, the expected ratio r is*

$$r = \frac{|\mathbb{A}|}{|\mathbb{D}|} \tag{1}$$

Definition 4 Positive Cluster: *A cluster is termed as positive, if the ratio of triangles from conformations of active molecules (active triangle) is significantly more than the expected ratio. Mathematically, cluster $\mathbb{C}$ is positive if it satisfies the following condition*

$$\mathbb{C} \textit{ is positive} : \frac{|\mathbb{P}|}{|\mathbb{C}|} \geq \delta r \tag{2}$$

where δ is a user-defined threshold parameter.

Definition 5 Negative cluster: *A cluster is termed negative, if the ratio of triangles from active class is significantly less than the expected ratio. Mathematically*

$$\mathbb{C} \textit{ is negative} : \frac{|\mathbb{P}|}{|\mathbb{C}|} \leq \frac{r}{\delta} \tag{3}$$

If a cluster is positive or negative, then it is a subspace of interest. More specifically, a positive cluster contains triangles that have a higher chance of binding to the target than triangles in other clusters. On the other hand, a negative cluster has a high concentration of triangles originating from inactives. As a result, triangles in a negative cluster are likely to be incompatible with the binding site in the target. In general, since the distribution of active triangles in a significant subspace deviates from the expected behavior, they have more discriminative power. As a result, positive and negative clusters can be employed to better understand the binding behavior of molecules and applied in higher level querying and mining tasks. For example, the triangle representing the cluster center can be employed to search molecules that display a desired activity. We show one such application in Molecular Classification in the next section.

Chart 1

```
Algorithm 1 $k$-medoid clustering
Require: $\mathbb{T}$ is a set of triangles of the same type
Require: $k$ is the desired number of clusters
Ensure: returns $k$ clusters
  $\mathbb{K} \leftarrow k$ randomly chosen triangles as cluster centers
  repeat
    $\mathbb{C}_i \leftarrow \emptyset, \forall i \in \{1 \cdots |\mathbb{K}|\}$
    for each triangle t $\in \mathbb{T} \backslash \mathbb{K}$ do
      $t \in \mathbb{C}_i$, where $rmsd(t, \mathbb{K}_i) \leq rmsd(t, \mathbb{K}_j)$, $j \neq i$
    for each cluster of triangles $\mathbb{C}_i$, where $\mathbb{C}_i \subset \mathbb{T}$ do
      $\mathbb{K}_i \leftarrow t$, where $t \in \mathbb{C}_i$, $\sum rmsd(t, t_n) \leq \sum rmsd(t_m, t_n) \forall t_m, t_n \in \mathbb{C}_i$, $t_n \neq t_m \neq t$
  until $\mathbb{K}$ has converged
  return $\mathbb{C}$
```

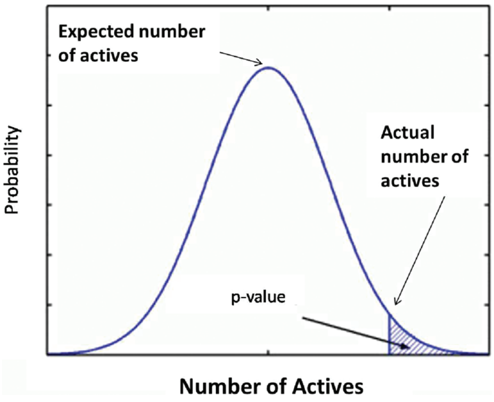


Figure 5. Measuring *p*-value from the probability distribution function.

Besides the identification of positive and negative clusters, an extensive significance analysis can also be performed using the idea of *p-value*.

Definition 6 *p*-VALUE: *The p-value of a cluster* $\mathbb{C}$ *that contains the set of active triangles* $\mathbb{P}$ *is defined as the probability that a random cluster of triangles contains more than* $|\mathbb{P}|$ *active triangles.*

As shown in Figure 5, the *p*-value of a cluster can be calculated by measuring the area to the right of the actual number of actives under its *probability distribution function (pdf)*. Clearly, the lower the *p*-value of a cluster, the more significant it is. The distribution of active triangles in a cluster can be modeled as a binomial. Each triangle in a cluster can be viewed as a trial, and a triangle being active can be regarded as "success". A cluster containing *m* triangles will involve *m* trials. The number of active triangles in the cluster is the number of successes. Therefore, the probability of a cluster $\mathbb{C}$ having μ active triangles is

$$P(\mathbb{C}; \mu) = \binom{m}{\mu} r^{\mu} (1-r)^{m-\mu} \tag{4}$$

where *r* is the expected ratio computed using eq 1, and $m = |\mathbb{C}|$.

The pdf of $\mathbb{C}$ can be generated from eq 4 by varying μ in the range $[0,m]$. Therefore, given the actual number of active triangles $\mu_0 = |\mathbb{P}|$ in $\mathbb{C}$, its *p*-value can be calculated by measuring the area under the pdf in the range $[\mu_0, m]$, which is

$$p\text{-}value(\mathbb{C}, \mu_0) = \sum_{i=\mu_0}^{m} P(\mathbb{C}; i) \tag{5}$$

Equation 5 reduces to the regularized Beta function $I(P(\mathbb{C}); \mu_0, m)$,[39] which is faster to compute. The *p*-value of a negative cluster can be computed using the same framework, except for the change of using an expected ratio of inactives rather than expected ratio of actives.

To determine statistically significant clusters, a threshold α is selected, and any cluster with a *p*-value less than α is considered statistically significant. Typically, α is set to 0.05. However, with multiple clusters, triangles from actives (or inactives) can get grouped together just by chance. Thus, to accurately classify a cluster as significant, we need to consider the factor of spurious positives resulting from multiple comparisons. Bonferroni correction addresses this issue. The Bonferroni correction is a multicomparison correction that is used when several dependent or independent statistical tests are being performed simultaneously. More specifically, if one wants a significance threshold α for the whole family of *n* tests, then Bonferroni correction lowers the significance threshold to α/n for each of the individual tests. For example, when the pharmacophore space is clustered into 50 clusters, for a cluster to be significant, its *p*-value needs to be less than 0.05/50.

Molecular Classification. In this section, we demonstrate the application of mining the joint pharmacophore space in molecular classification. Pharmacophores corresponding to significant cluster centers are indicative of the selective chemical and biological activity. This can be used for *in silico* prediction of activity of new molecules through a classification approach.

Figure 6 presents the workflow of the classification algorithm. Given a training data set with molecules labeled as active or inactive, we first cluster the joint pharmacophore space as described in the previous section. Next, we identify all *significant triangles* in this joint pharmacophore space and use them as pharmacophoric keys.

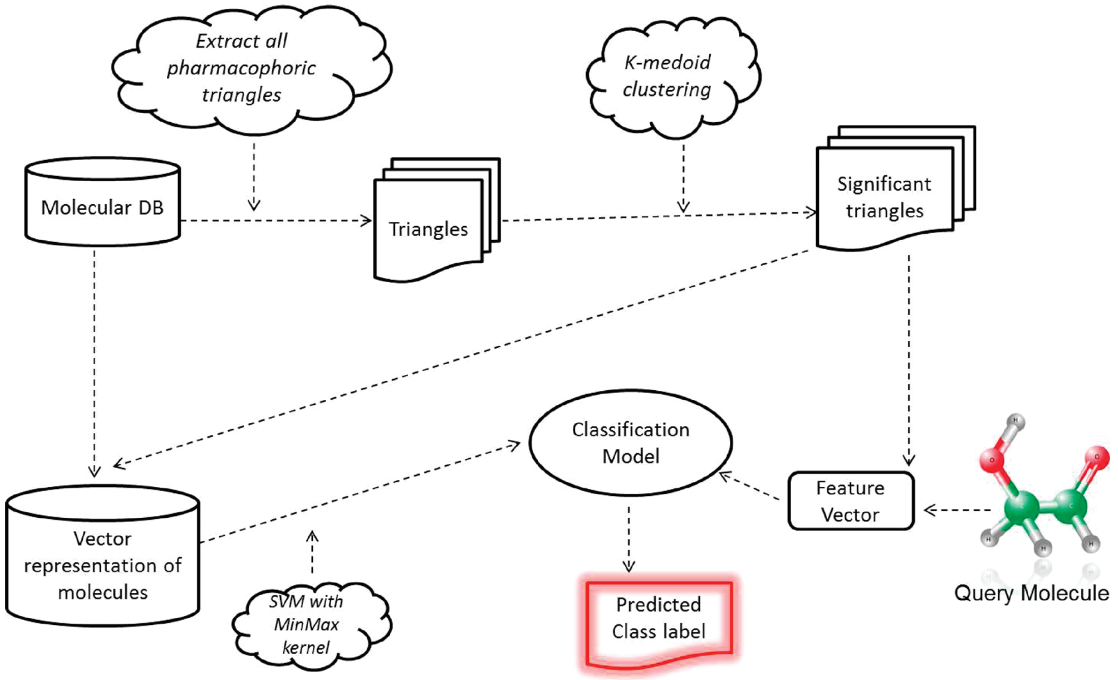


Figure 6. Outline of the proposed classification algorithm to classify molecules based on significant subspaces mined from the joint pharmacophore space.

Definition 7 Significant Triangle: *A triangle is termed significant if it forms the cluster center of a positive or negative cluster.*

The rationale behind using significant triangles as pharmacophoric keys is to characterize molecules based on how closely they align to the discriminative subspaces within the joint pharmacophore space. Since each of the positive or negative clusters deviates from the expected behavior, their centers have a high discriminative power and provide an excellent platform to build classifiers.

Once the significant triangles are identified, all molecules in the training data set are converted to a feature vector where each dimension corresponds to a specific significant triangle. Algorithm 2 (see Chart 2) presents the pseudocode to convert a given molecule to a feature vector. Given a molecule *m*, first, all triangles in *m* are identified (line 1). Next, each of the extracted triangles is compared to the significant triangles to identify the closest significant triangle (lines 3−15). If the rmsd of the closest significant triangle is within a user-specified threshold, then the dimension corresponding to the significant triangle is incremented (lines 14−15). Essentially, for each triangle in the query molecule, we check whether it aligns well with any of the significant triangles. Based on this result, the information is stored in the vector representation of the molecule. Ultimately, the vector representation of the query molecule is returned (line 16).

Given the vector representation of the molecule in the training data set, we use support vector machines (SVM)[40] to develop the training model. One key issue that affects the performance of SVM is the choice of kernel. The kernel function computes the similarity between two input vectors. Theoretically, any kernel can be used as long as the similarity matrix computed by the kernel function satisfies the Mercer's conditions.[41] We use the MinMax kernel to build the classification model. The MinMax kernel function for vectors $X = [x_1,\cdots,x_n]$ and $Y = [y_1,\cdots,y_n]$ is defined as follows

$$K(X, Y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)} \; \forall i \qquad (6)$$

The MinMax kernel function is similar to the Tanimoto coefficient, which has been extensively used in the chemoinformatics community.[42−45] For binary vectors, both kernels return the same similarity value. However, since our vectors contain actual counts, we use the MinMax kernel. The MinMax kernel has been shown to satisfy the Mercer's conditions.[41]

Once the classification model is built, a query molecule (specifically, its 3D conformation) is converted to a feature vector using the same procedure in Algorithm 2 (see Chart 2). Next, the vector is provided as input to the classifier to obtain its predicted class.

■ RESULTS

Data Sets. For a thorough evaluation, we benchmark the proposed techniques on 32 data sets obtained from three different sources.

1 We use 11 anticancer screen data sets from the PubChem[46] data repository. PubChem is a well-maintained compilation of biological activities of various molecules, containing bioassay records for anticancer screen data sets against various cancer cell lines. Each data set contains molecules tested against a particular cancer cell line and the outcome of the test: *active* or *inactive*. A brief summary of each of these NCI bioassays is provided in Table 1.

2 We also use a data set of 10,000 molecules that were assayed against cyclin-dependent kinase 5 (CDK-5) to identify CDK-5 inhibitors. CDK-5 is widely viewed as a possible

Chart 2

Algorithm 2 Conversion to Feature Vector($\mathbb{S}, m$)

```
Require: $\mathbb{S}$ is the set of significant triangles mined from the training dataset
Require: $m$ is the query molecule
Require: $\theta$ is the minimum distance threshold
Ensure: returns vector representation of $m$
$\mathbb{T} \leftarrow$ all triangles in $m$
$\underline{V} \leftarrow$ vector of size $|\mathbb{S}|$ initialized to 0
for each $t \in \mathbb{T}$ do
    dim $\leftarrow 1$
    minDist $\leftarrow \infty$
    minDim $\leftarrow 0$
    for each $s \in \mathbb{S}$ do
        if $t$ is mappable to $s$ then
            $d \leftarrow \mathrm{rmsd}(t, s)$
            if $d <$ minDist then
                minDist $\leftarrow d$
                minDim $\leftarrow$ dim
        dim $\leftarrow$ dim+1
    if minDist $< \theta$ then
        $V$[minDim]$\leftarrow V$[minDim]+1
return $\underline{V}$
```

Table 1. Anticancer Screen Data Sets

name	size	number of actives	description
MCF-7	22942	1989	breast
MOLT-4	32902	1970	leukemia
NCI-H23	33160	1178	non-small cell lung
OVCAR-8	33247	1219	ovarian
P388	33796	1182	leukemia
PC-3	22672	942	prostate
SF-295	33080	1203	central nervous system
SN12C	32794	1114	renal
SW-620	33283	1419	colon
UACC-257	32870	923	melanoma
yeast	64110	6562	yeast anticancer

target for a wide variety of neurological disorders. Out of the 10,000 molecules assayed, only 102 were found to be active.

3 The third set contains data sets obtained from the DUD repository.[47] DUD is a directory of useful "decoys" for benchmarking virtual screening techniques. More specifically, DUD contains active molecules against a series of targets. For each active, the directory also contains 36 "decoys" with similar physical properties (such as molecular weight, calculated LogP) but dissimilar topology. Due to this unique selection of decoys, the DUD data set has been shown to be harder to classify than "uncorrected" data sets such as the MDDR.[47] For benchmarking the proposed techniques, we choose actives and their corresponding decoys against 20 different targets. Table 2 provides a brief summary of each of these data sets.

Experimental Setup. For each molecule in the cancer data sets, we use a low-energy conformer generated by PubChem. The conformers are generated from a conformer model describing energetically accessible and biologically relevant conformations of the compound structure. For the CDK-5 and DUD data sets, the conformers are generated using the Joelib computational library.[48] The pharmacophoric features used are hydrogen bond donors and acceptors, aromatic rings, hydrophobic cores (*abbrv.* hyd), and groups with positive and negative charges (*abbrv.* cation and anion, respectively). All experiments were performed on a machine with a 3.2 GHz Intel Xeon processor, 4GB memory, and running Debian Linux 4.0. All our algorithms are implemented in Java 1.6.0.

Significant Subspaces in the Joint Pharmacophore Space. In this section, we analyze the joint pharmacophore space of various data sets and mine statistically significant subspaces under three different circumstances: actives and inactives against a single target, actives and inactives from a cell-based assay with potentially multiple unknown targets, and actives and inactives against multiple known targets. For each of these cases, we present a subset of the significant subspaces to demonstrate the utility of the proposed technique. The significant subspaces indicate specific geometric patterns that correlate with a biological activity. As a result, they serve as an excellent platform to develop further molecular analysis tools. We specifically analyze

Table 2. DUD Data Sets

target	size	number of actives	description
ACE	1846	49	angiotensin-converting enzyme
ACH	3999	107	acetylcholine esterase
ALR2	1021	26	aldose reductase
AmpC	807	21	AmpC beta lactamase
AR	2933	79	androgen receptor
CDK2	2146	72	cyclin dependent kinase 2
COX-2	13715	426	cyclooxygenase 2
DHFR	8777	410	dihydrofolate reductase
EGFr	16471	475	epidermal growth factor receptor kinase
FXa	5891	146	factor Xa
GPB	2192	52	glycogen phosphorylase beta
HMGR	1515	35	hydroxymethylglutaryl-CoA reductase
NA	1923	49	neuraminidase
P38	9595	454	P38 mitogen activated protein kinase
PARP	1386	35	poly(ADP-ribose) polymerase
PDGFrb	6150	170	platlet derived growth factor receptor kinase
SRC	6478	159	tyrosine kinase SRC
thrombin	2528	72	thrombin
TK	913	22	thymidine kinase
VEGFr2	2994	88	vascular endothelial growth factor receptor kinase

Table 3. Positive Significant Subspaces in the Joint Pharmacophore Space of the CDK-5 Data Set

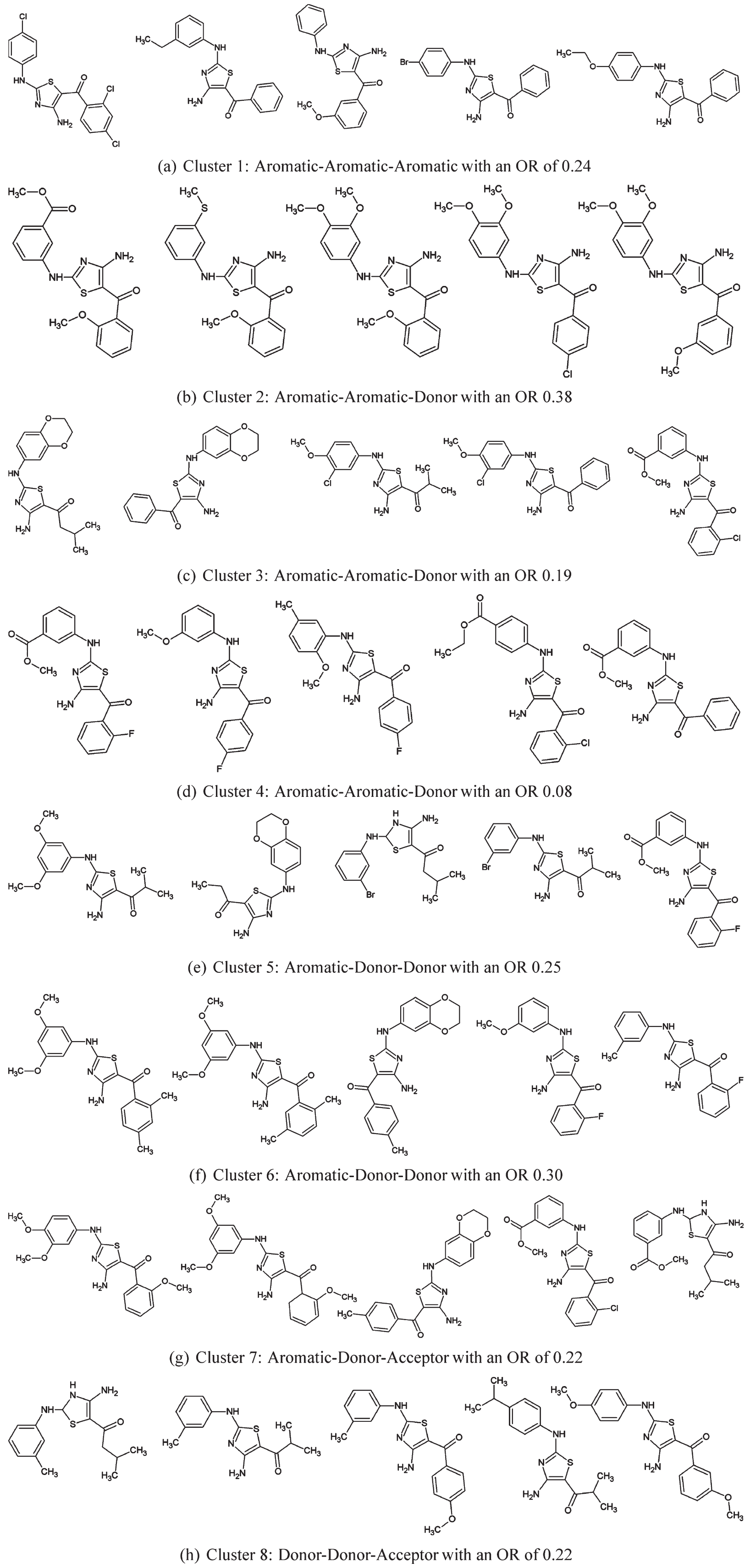
cluster ID	triangle type	cluster size	OR	*p*-value
1	aromatic−aromatic−aromatic	268	0.24	1.75×10^{-68}
2	aromatic−aromatic−donor	455	0.38	1.71×10^{-223}
3	aromatic−aromatic−donor	545	0.19	5.54×10^{-94}
4	aromatic−aromatic−donor	625	0.08	1.31×10^{-31}
5	aromatic−donor−donor	436	0.25	4.47×10^{-123}
6	aromatic−donor−donor	353	0.3	4.56×10^{-118}
7	aromatic−donor−acceptor	461	0.22	6.38×10^{-102}
8	donor−donor−acceptor	469	0.22	6.92×10^{-104}

their performance in molecular classification in the next section. To identify positive and negative clusters, we set the value of δ in eq 2 and eq 3 to 2 and k to 50. The value of k is selected by employing the "elbow method".[49] Any of the other methods to determine the number of clusters[50] can be applied as well. The choice of δ is not as critical to the mining procedure as k since the parameter simply defines the strictness of the positive/negative filtering criteria. The smaller the value of δ, the larger is the answer set. For a focused study of the significant subspaces, a larger δ is optimal, whereas to obtain a general perspective of the distribution of the subspaces, a smaller δ is more appropriate.

Actives and Inactives against a Single Target. Table 3 presents a subset of the positive significant subspaces mined from the joint pharmacophore space of the CDK-5 database where only 1% of the molecules are active. Figure 7 presents some representative molecules containing triangles from each of the clusters shown in Table 3. We only present positive clusters, since most of the remaining clusters are negative. The large number of negative clusters is a direct consequence of triangles from conformations of active molecules getting clustered together. The subspaces are mined using the algorithm for *k*-medoid clustering described earlier. As can be seen, all the subspaces shown in Table 3 have an observed ratio (*abbrv.* OR) much higher than the expected ratio of 0.01 resulting in extremely low *p*-values. Applying Bonferroni correction at a significance threshold of 0.05 on 50 clusters, any cluster with a *p*-value lower than 0.001 is statistically significant.

An illustration of the above event can be seen in the triangle type <aromatic−aromatic−aromatic>. Among all the 50 clusters in this triangle type, the single cluster shown in Table 3 (Cluster ID 1) contains all triangles originating from conformations of active molecules. As a result, all remaining clusters in this triangle type are negative. This is a highly significant event, since it precisely highlights the geometric configuration of aromatic ring centers that is essential for molecules to be CDK-5 inhibitors. A similar behavior is also observed in the triangle group <aromatic−donor−donor>, where triangles from actives are distributed only among six clusters including the three shown in Table 3. Overall, these results indicate our technique to be successful in identifying subspaces that are more likely to contain active triangles in the joint pharmacophore space of the CDK-5 data set.

Actives and Inactives from Cell-Based Assays. To analyze the performance of the proposed technique on cell-based assays, we use the cancer data sets. The unique property of cell-based assays is that a molecule is not labeled based on whether it binds to a specific target. Rather, a molecule is judged based on whether it is having the desired effect on a cell line. The desired effect can in the result of a molecule binding to any one target from a set of potentially unknown targets. Thus, all actives cannot be assumed to be binding to the same target. Consequently, techniques that optimize pharmacophores against a single target cannot be applied in this setting. On the other hand, our proposed technique can identify the geometries that correlate with binding activity even in the multiple-unknown-targets setting by clustering the joint pharmacophore space and analyzing their statistical significances.



(a) Cluster 1: Aromatic-Aromatic-Aromatic with an OR of 0.24

(b) Cluster 2: Aromatic-Aromatic-Donor with an OR 0.38

(c) Cluster 3: Aromatic-Aromatic-Donor with an OR 0.19

(d) Cluster 4: Aromatic-Aromatic-Donor with an OR 0.08

(e) Cluster 5: Aromatic-Donor-Donor with an OR 0.25

(f) Cluster 6: Aromatic-Donor-Donor with an OR 0.30

(g) Cluster 7: Aromatic-Donor-Acceptor with an OR of 0.22

(h) Cluster 8: Donor-Donor-Acceptor with an OR of 0.22

Figure 7. Representative molecules from the positive significant subspaces of the CDK-5 data set.

Table 4. Positive Significant Subspaces in the Joint Pharmacophore Space of the Cancer Data Sets

data set	triangle type	cluster size	ER	OR	*p*-value
OVCAR-8	acceptor—cation—cation	284	0.037	0.246	1×10^{-37}
OVCAR-8	acceptor—aromatic—cation	207	0.037	0.25	3.86×10^{-29}
OVCAR-8	acceptor—donor—donor	1283	0.037	0.125	1.71×10^{-40}
OVCAR-8	aromatic—donor—donor	853	0.037	0.158	5.44×10^{-46}
OVCAR-8	acceptor—acceptor—acceptor	3886	0.037	0.135	8.55×10^{-141}
P388	acceptor—acceptor—donor	1932	0.035	0.226	4.04×10^{-213}
P388	acceptor—acceptor—cation	9832	0.035	0.21	1.1×10^{-96}
P388	acceptor—aromatic—Hyd	983	0.035	0.133	5.37×10^{-39}
PC-3	acceptor—cation—cation	133	0.041	0.316	9.38×10^{-27}
PC-3	acceptor—cation—donor	216	0.041	0.282	1.79×10^{-34}
PC-3	acceptor—donor—donor	2852	0.041	0.134	2.05×10^{-90}
PC-3	acceptor—donor—aromatic	5361	0.041	0.123	8.5×10^{-136}
SF-295	acceptor—cation—cation	265	0.036	0.25	9.85×10^{-37}
SF-295	cation-donor—donor	330	0.036	0.245	6×10^{-44}
SF-295	acceptor—acceptor—donor	4011	0.036	0.15	2.85×10^{-195}
SF-295	aromatic—aromatic—donor	543	0.036	0.225	3.81×10^{-60}
SW-620	acceptor—cation—cation	226	0.0426	0.265	2.2×10^{-31}
SW-620	cation-donor—donor	206	0.0426	0.257	2.74×10^{-27}
SW-620	aromatic—donor—donor	4260	0.0426	0.13	1.13×10^{-116}
SW-620	acceptor—donor—donor	3578	0.0426	0.137	2.51×10^{-113}
SW-620	acceptor—donor—Hyd	687	0.0426	0.21	4.44×10^{-58}

Table 4 presents a subset of the positive clusters from the cancer data sets. Similar to the results on the CDK-5 data set, the subspaces shown in Table 4 have a ratio of active triangles that is much higher than the expected. The low *p*-values of these subspaces quantify their statistical significance. An interesting pattern is also observed in the triangle type <acceptor—acceptor—anion> across multiple cancer data sets. There is a recurring negative cluster in this triangle type across the cancer data sets. This clustering pattern highlights a global negative binding affinity for this triangle type.

Actives and Inactives against Multiple Known Targets. To prepare a data set of actives against multiple known targets, we utilize the DUD data sets. Specifically, we prepare a data set by merging the ligands and decoys for four different targets: DHFR, EGFr, PARP, and PDGFrb. As mentioned earlier, the DUD data set provides an excellent benchmarking platform since the decoys have been selectively added to the repository due to their physical similarity to the ligands. To show the utility of the proposed technique, we identify the positive clusters in the joint pharmacophore space of the data set and demonstrate how each such cluster correlates with the four different targets.

Table 5 presents the positive clusters corresponding to each of the targets in the multitarget setting. The correspondence between a target and a cluster is inferred by computing the target-type of the majority of the actives in the cluster. As can be seen, the OR of each of the clusters in Table 5 is significantly higher than its expected ratio (*ER*). The best example of the phenomenon is Cluster 5, where the OR is 0.69 even though the ER is 0.012. Furthermore, the extremely low *p*-values of the clusters substantiate their statistical significances. This result shows that even in the multitarget setting where actives from multiple targets are mixed with inactives, there are regions in the joint pharmacophoric space that are statistically significant due to high concentration of triangles originating from the actives. Since the triangles corresponding to the binding site of the active molecules of a specific target are geometrically similar, they colocate the same region in the pharmacophoric space. The proposed technique is able to identify those subspaces by analyzing their statistical significances and correlate the geometries to binding activity.

An important question that arises from the above analysis is whether each positive cluster corresponds to a single target. To answer this question, we inspect the distribution of the four target-types of the actives in the positive clusters. Table 6 presents the distribution. As can be seen, most of the positive clusters are dominated by actives corresponding to a single target. For example, clusters 5, 10, and 14 contain actives of only DHFR, EGFr, and PARP, respectively. There are few positive clusters, such as clusters 13 and 17, that contain significant number of actives from two different targets. Such a case is possible if the geometries required for ligands of two targets share some commonalities. Thus, unless there is topological similarity between two targets, each positive cluster is likely to correspond to a single target. Overall, the result establishes that each positive cluster typically describes the geometric configuration necessary for ligands to bind to a specific target.

While Table 6 shows that each positive cluster typically relates to a single target, a question arises about the interpretation of multiples positive clusters mapping to a single target. To gain an intuition behind this behavior, recall that each cluster is comprised of pharmacophoric triangles extracted from the database molecules. At the same time, the geometry responsible for the binding between a ligand and the target may be larger than just a triplet. Thus, the binding geometry can be decomposed into multiple triangles. As a result, each such triangle is likely to correspond to a positive cluster mapping to the same target. Following the same reasoning, when a positive cluster contains actives from multiple targets, it is likely that the representative triangle of the cluster is shared by the geometries responsible for binding to each of the targets.

Table 5. Positive Significant Subspaces in the Joint Pharmacophore Space of the DUD Data Sets

cluster ID	target	triangle type	cluster size	ER	OR	*p*-value
1	DHFR	aromatic–aromatic–aromatic	468	0.012	0.34	8.39×10^{-184}
2	DHFR	aromatic–aromatic–donor	1965	0.012	0.16	1.62×10^{-226}
3	DHFR	aromatic–acceptor–acceptor	4645	0.012	0.12	0
4	DHFR	aromatic–acceptor–acceptor	7856	0.012	0.12	0
5	DHFR	aromatic–donor–donor	508	0.012	0.69	0
6	DHFR	acceptor–acceptor–donor	1136	0.012	0.11	1.67×10^{-73}
7	DHFR	acceptor–donor–donor	884	0.012	0.23	1.28×10^{-191}
8	DHFR	acceptor–donor–donor	1840	0.012	0.27	0
9	EGFr	aromatic–aromatic–aromatic	1089	0.014	0.33	0
10	EGFr	aromatic–aromatic–aromatic	132	0.014	0.26	3.88×10^{-34}
11	EGFr	aromatic–aromatic–acceptor	2809	0.014	0.15	2.82×10^{-295}
12	EGFr	aromatic–aromatic–donor	1049	0.014	0.24	9.49×10^{-218}
13	EGFr	aromatic–donor–donor	614	0.014	0.16	9.17×10^{-69}
14	PARP	aromatic–aromatic–acceptor	2051	0.001	0.04	2.69×10^{-83}
15	PDGFrb	aromatic–aromatic–aromatic	311	0.005	0.31	7.77×10^{-144}
16	PDGFrb	aromatic–aromatic–donor	445	0.005	0.17	1.35×10^{-92}
17	PDGFrb	aromatic–aromatic–donor	817	0.005	0.15	1.92×10^{-139}
18	PDGFrb	aromatic–donor–donor	266	0.005	0.25	2.26×10^{-91}
19	PDGFrb	acceptor–donor–donor	533	0.005	0.12	7.2×10^{-64}

Table 6. Distribution of the Target-Types in the Positive Clusters of DUD Data Set

cluster ID	DHFR	EGFr	PARP	PDGFrb
1	**0.34**	0.01	0	0
2	**0.16**	0.02	0	0.04
3	**0.12**	0	0	0
4	**0.12**	0.04	0	0
5	**0.69**	0	0	0
6	**0.11**	0.01	0	0
7	**0.23**	0.03	0	0
8	**0.27**	0.01	0	0.05
9	0.05	**0.33**	0	0
10	0	**0.26**	0	0
11	0.04	**0.15**	0	0.02
12	0	**0.24**	0	0
13	0.10	**0.16**	0	0
14	0	0	**0.04**	0
15	0	0.04	0	**0.31**
16	0.03	0.03	0	**0.17**
17	0.13	0.06	0	**0.15**
18	0	0.03	0	**0.25**
19	0	0.02	0	**0.12**

The presented results highlight the advantage of the proposed technique over single-target based approaches. A number of key conclusions can be drawn from the above results. First, as shown in Table 4 and Table 5, even when ligands of topologically dissimilar targets are mixed together, the proposed algorithm is able to identify regions in the pharmacophoric space that are statistically significant. Second, as shown in Table 6, the proposed algorithm can further correlate each of the significant subspaces to a specific target thereby, eliminating the need to perform pharmacophore analysis on a target-by-target basis.

The proposed technique can be applied in a number of ways. In data sets such as cell-based assays where the targets may not be known, the actives can be automatically categorized based on the subspaces to which they conform. Each of these subspaces can then be employed for more sophisticated querying and mining tasks. Significant triangles can also be employed to quantify molecules based on how well they conform to a desired pharmacophore model. For example, given a set of triangles from positive clusters, and another set of triangles representing negative clusters, molecules in a database can be scored based on their similarity to the triangles. Such a score would reward database molecules if they contain triangles that are geometrically similar to the positive triangles and penalize molecules if they contain triangles that are structurally similar to negative triangles. This kind of substructure queries based on significant triangles is not possible in techniques such as fingerprints or 3-point pharmacophores since they compute a global description of the molecules. Furthermore, fingerprinting techniques are typically exhaustive in nature and employ all features in their specific domain to characterize a molecule. As a result, they cannot be used to query molecules on a specific set of features that are more important for a desired binding activity to take place.

The proposed technique can also be employed to check if the binding sites of two targets are topologically similar. The similarity in the binding sites of two targets can be captured by finding the correlation coefficient between the positive clusters of the pharmacophore space. More specifically, when the binding sites of two targets are topologically similar, the corresponding actives will have pharmacophoric triangles that are geometrically similar as well. Consequently, the triangles are likely to display membership in the same positive clusters. This comembership can be captured by computing the correlation between the target-type distributions of the actives among the positive clusters.

Since each of the significant subspaces are discriminative in nature, their representative triangles can also be used for molecular classification. In the next section, we analyze this application in more detail. Benchmarking the potential of significant triangles in molecular classification also allows us to correlate them with chemical properties. More specifically, if the significant triangles are not indicative of the chemically properties, then their performance in classification is likely to be, at best, average.

Classification Performance. To demonstrate the utility of the significant subspaces, we use representative triangles from each of these subspaces and use them as keys to build a molecular descriptor (*abbrv.* JPS-d). For benchmarking the proposed descriptor, we compare the performance with descriptors from three other approaches:

1. **Pharmacophoric keys:** For pharmacophoric keys, we use the Joelib2 computational chemistry library[48] to generate the bit vector representation of a molecule. Joelib generates a 54-bit presence/absence feature vector where each bit corresponds to a structural key. The structural keys range from certain functional groups, to the presence of pharmacophoric features such as hydrogen bond donors and acceptors.
2. **3-Point pharmacophores:** 3-Point pharmacophores (*abbrv.* 3pp)[34,35] is based on the idea of enumerating all triplets in a molecule and then hashing them to generate molecular fingerprints. Theoretically, 3pp is the closest to our approach among existing techniques in the chemoinformatics literature. Both approaches use the same primitive features. However, the process of going from this feature space to the vector representation is different.

   In 3pp, the pharmacophore space is defined by all combinations of 3-point pharmacophores, together with all distance ranges between each of the triangle vertices. The feature space is binned by dividing the interfeature distances into ranges. Due to this binning, two triplets will fall into the same bin in the pharmacophore space, if they are mappable to each other and the interfeature distances of the triplets are within the exact same ranges. To get a better intuition of the binning in the pharmacophore space, consider this numerical example. If there are six pharmacophoric features, and the interfeature distances are divided into six bins, then the total number of bins in the pharmacophore space will be

   $$\left(\binom{6}{3} + \binom{6}{2} \times 2 + \binom{6}{1}\right) \times 6^3 = 12096$$

   However, a number of these bins correspond to triangles that are geometrically infeasible due to the triangle inequality rule. The infeasible bins are removed to keep the space compact.

   The vector representation of a molecule reflects the distribution of the triplets in the pharmacophore space. More specifically, the number of dimensions in the vector is equal to the number of feasible bins in the pharmacophore space. A dimension value is set to 1 if the corresponding bin in the pharmacophore space contains at least one triplet, otherwise it is set to 0.

   A number of techniques exist to generate 3pp from molecules.[34,35] While the underlying methodology is similar in all of the existing methods, differences exist in the number of pharmacophoric features extracted, and the division of the interfeature ranges into bins. Similar to 3pp, 4-point pharmacophores exist as well. However, 4-point pharmacophores suffer from the combinatorial explosion in the size of the pharmacophore space. To benchmark our approach, we choose the specifications mentioned in Loob fingerprints.[34] Loob extracts the same six pharmacophoric features used in our approach. The interfeature distances are binned into the following six ranges: 0−4.5 Å, 4.5−7 Å, 7−10 Å, 10−14 Å, 14−19 Å, 19−24 Å.
3. **Fingerprints:** For fingerprints, we choose Daylight[33] and Molprint2D,[51] two of the most well-known fingerprinting techniques in the chemoinformatics community. For Daylight, we use the default parameters to generate a 2048 bit binary feature vector where all paths of length 7 in a molecule are enumerated and then hashed to generate its vector representation. In Molprint2D, each bit in the fingerprint is derived from a tabular data structure that stores the central heavy atom-type and a list of all other heavy atom-types within a distance of two bonds.

To make a uniform comparison, we use SVM as the classification algorithm for each of the above methods using the libsvm[52] tool. The classification performance is evaluated by performing 5-fold cross-validation on molecules sampled from each of the cancer and DUD data sets. More specifically, we sample up to 1500 molecules from each of the cancer and DUD data sets. For data sets containing less than 1500 molecules, we use the entire data set. In each of the sampled data sets, we keep the distribution actives and inactives uniform, unless a data set contains less than 750 actives. In such a case, we include the entire set of actives in the sampled set. Next, we divide each of the sampled data sets into five subsets and then four of these subsets are merged to form the training set, while the fifth subset is retained as the testing set for validation. The process is repeated five times by using each of the five subsets once as the testing set. Moreover, the cross-validation process is repeated five times using five randomly drawn samples from each of the cancer and DUD data sets to measure consistency in performance. The MinMax kernel is used for JPS-d, whereas the Tanimoto kernel is used for the rest of the descriptors.

For our approach, the pharmacophoric keys are mined by only analyzing the molecules in the training set for a particular fold. To identify positive and negative clusters, we set the value of δ to 1.5. Convergence of cluster centers is determined using a distance threshold of 0.2 Å. More specifically, the clustering process continues until the distance between the cluster centers in two consecutive iterations is less than the specified threshold. The cluster centers are then used as keys to convert molecules in both the training and testing set into feature vectors. Vectors in the training data set along with their class labels are then analyzed to learn the classification model using SVM. Finally, the class labels are predicted for molecules in the testing set.

The classification quality on these data sets is quantified using two well-established metrics: ROC50 and BEDROC.[53] Receiving Operating Characteristics (ROC) curve is a graphical plot of the *true positive rate* against the *false positive rate* for a classifier as the discrimination threshold is varied. The area under the ROC curve (AUC) is a measure of the classification accuracy. The area is bounded within the range [0,1] where a perfect model will achieve an AUC of 1. ROC50 is a modification of the ROC measure to focus more on the early part of the ranking. More specifically, ROC50 computes the area under the ROC curve until 50 false positives. As a result, ROC50 is a preferred measure of performance in virtual screening of molecules where early recognition of actives is a desirable quality.

Table 7 presents the area under the ROC50 curves for each of the descriptors across all the cancer data sets. The best ROC50

Table 7. ROC50 Comparison between 3pp, Joelib, Daylight, JPS-d, and Molprint2D on the Cancer Data Sets

data set	3pp	Joelib	Daylight	Molprint2D	JPS-d
MCF-7	0.32	0.42	0.50	**0.52**	0.50
MOLT-4	0.50	0.44	0.41	0.49	**0.59**
NCI-H23	0.62	0.51	0.56	0.64	**0.68**
OVCAR-8	0.56	0.53	0.56	0.60	**0.62**
P388	0.55	0.45	0.59	**0.67**	0.57
PC-3	**0.60**	0.54	0.58	0.54	**0.60**
SF-295	0.62	0.51	0.61	0.62	**0.63**
SN12C	0.63	0.55	0.59	**0.64**	0.60
SW-620	0.51	0.53	0.51	**0.62**	0.56
UACC-25	0.58	0.51	0.57	0.59	**0.63**
Yeast	0.42	0.41	**0.45**	0.33	0.44

Table 8. ROC50 Comparison between 3pp, Joelib, Daylight, JPS-d, and Molprint2D on the DUD Data Sets

data set	3pp	Joelib	Daylight	Molprint2D	JPS-d
ACE	0.69	0.42	0.75	0.87	**0.92**
ACHE	0.74	0.52	0.83	0.88	**0.90**
ALR2	0.72	0.60	0.58	**0.94**	0.93
AmpC	0.69	0.68	**0.83**	0.78	0.81
AR	0.71	0.77	0.82	0.87	0.82
CDK2	0.77	0.55	0.75	0.89	0.76
COX-2	0.89	0.84	0.92	0.92	**0.94**
DHFR	0.88	0.84	**0.92**	0.83	**0.92**
EGFr	0.88	0.86	**0.97**	0.89	0.93
FXa	0.74	0.51	**0.91**	0.88	**0.91**
GPB	0.78	0.72	0.72	0.78	**0.94**
HMGR	0.89	0.88	0.79	**0.97**	0.88
NA	0.84	0.83	0.81	0.85	**0.87**
P38	0.95	0.85	**0.95**	0.90	**0.95**
PARP	0.85	0.83	0.82	0.89	0.99
PDGFrb	0.89	0.66	**0.94**	0.83	**0.94**
SRC	0.87	0.83	0.93	0.88	**0.94**
Thrombin	0.82	0.72	0.76	**0.87**	0.82
TK	0.42	0.43	0.81	0.54	**0.87**
VEGFr2	0.60	0.41	0.79	**0.89**	0.82

Table 9. BEDROC Comparison with Daylight, 3pp, Joelib, and Molprint2D on the Cancer Data Sets

data set	3pp	Joelib	Daylight	Molprint2D	JPS-d
MCF-7	0.28	0.28	0.41	0.43	**0.58**
MOLT-4	0.44	0.35	0.42	0.51	**0.54**
NCI-H23	0.44	0.28	0.44	0.51	**0.57**
OVCAR-8	0.50	0.30	0.40	0.63	**0.68**
P388	0.23	0.37	**0.50**	0.44	**0.50**
PC-3	**0.66**	0.46	0.33	0.49	0.60
SF-295	0.56	0.38	0.32	0.56	**0.63**
SN12C	0.54	0.42	0.40	0.70	0.67
SW-620	0.42	0.36	0.36	0.31	**0.67**
UACC-257	0.37	0.34	0.34	0.48	**0.68**
Yeast	**0.38**	0.20	**0.38**	0.28	**0.38**

Table 10. BEDROC Comparison between 3pp, Joelib, Daylight, JPS-d, and Molprint2D on the DUD Data Sets

data set	3pp	Joelib	Daylight	Molprint2D	JPS-d
ACE	0.60	0.46	**0.91**	**0.91**	**0.91**
ACHE	0.65	0.42	0.94	0.93	**0.95**
ALR2	0.80	0.60	**0.88**	0.83	**0.88**
AmpC	0.75	0.66	0.98	0.95	**0.99**
AR	0.59	0.69	0.89	**0.97**	0.92
CDK2	0.68	0.46	0.93	**0.99**	0.94
COX-2	0.91	0.79	0.99	0.94	**0.99**
DHFR	**1**	0.92	**1**	**1**	**1**
EGFr	0.96	0.87	0.98	0.98	**0.99**
FXa	0.77	0.48	0.97	0.98	**0.98**
GPB	0.82	0.78	0.87	0.83	**0.99**
HMGR	0.90	0.95	0.86	0.96	0.87
NA	0.75	0.75	0.98	0.94	**0.99**
P38	**1**	0.77	**1**	**1**	**1**
PARP	0.88	0.84	**1**	**1**	**1**
PDGFrb	0.94	0.66	**1**	**1**	**1**
SRC	0.91	0.87	**1**	**1**	**1**
Thrombin	0.85	0.69	0.95	**0.99**	0.96
TK	0.73	0.26	0.85	**0.91**	0.86
VEGFr2	0.61	0.35	0.86	0.90	**0.92**

score for each data set is highlighted in bold. Clearly, the proposed descriptor outperforms 3pp, Joelib, and Daylight by a significant margin. Molprint2D achieves the highest ROC50 score on four data sets, whereas JPS-d achieves the highest score on six out of the 11 data sets. We observed that when the number of clusters identified as statistically significant is low, the classification performance suffers as well. This is particularly the case with the P388 and yeast data sets. In both of these, triangles from actives and inactives tend to colocate. As a result, the geometries of actives and inactives cannot be correlated with binding activity properly.

Table 8 presents the area under the ROC50 curve for each of the descriptors across the DUD data sets. Similar to the results observed on the cancer data sets, JPS-d achieves the best ROC50 score on most of the DUD data sets. More specifically, JPS-d performs at least as well or better in 12 out of the 20 data sets. The next best performance is achieved by Molprint2D which achieves the highest score in six data sets. Compared to the cancer data sets, the ROC50 scores are higher in the DUD data sets. This improvement is likely due to the fact that in the cancer data sets the actives are potentially binding to multiple targets.

To get a deeper understanding of the performance of JPS-d as a molecular descriptor, we also evaluate the classification quality on the Boltzmann-Enhanced Discrimination of Receiver Operator Characteristics (*abbrv. BEDROC*)[53] metric. BEDROC was developed to overcome the shortcoming of ROC in that it did not give enough importance to early recognitions of actives. Compared to ROC (or ROC50), BEDROC is more sensitive to the "early recognition problem" in virtual screening of molecules by rewarding early ranking of actives in an exponential manner. Other techniques to reward early retrieval of actives exist as well.[54−56] For example, the CROC[54] metric rewards early enrichment by smoothly magnifying portions of interest in the ROC curve using an appropriate continuous and monotone function.

Table 11. Spearman's Rank Correlation Coefficient between the Descriptors

descriptor combination	NCI-H23	PC-3	SF-295
Daylight - JPS-d	0.52	0.65	0.64
Molprint2D - JPS-d	0.56	0.54	0.51
3pp - JPS-d	0.75	0.74	0.76
Joelib - JPS-d	0.48	0.50	0.41
Molprint2D - 3pp	0.48	0.44	0.47
Molprint2D - Joelib	0.51	0.53	0.54
Molprint2D - Daylight	0.63	0.63	0.60
3pp - Joelib	0.36	0.37	0.37
3pp - Daylight	0.50	0.57	0.40
Joelib - Daylight	0.42	0.52	0.51

We set the α parameter in BEDROC to 20 which means the first 8% of the ranked list contributes 80% of the BEDROC score. Since BEDROC is sensitive to the "saturation effect", we reduce the ratio of actives in the testing sets to 0.01. The saturation effect can significantly influence the BEDROC score for the same "constant true performance" when the ratio of actives vary in the testing sets. More specifically, if the ratio of actives R_a is significantly high such that $\alpha R_a \ll 1$, then the BEDROC metric is unable to properly distinguish between a good and a bad performance. A detailed analysis behind this behavior is available in the work by Truchon et al.[53]

Table 9 shows the BEDROC scores for 3pp, Joelib, Daylight, Molprint2D, and JPS-d. As a reference point for performance evaluation, a random classifier would achieve a BEDROC score of 0.05. As can be seen, all five descriptors achieve scores that are significantly higher than 0.05. JPS-d performs as well or better in nine out of the 11 cancer data sets. 3pp and Molprint2D achieve a higher BEDROC score in one data set each. Overall, JPS-d outperforms all other descriptors by a significant margin.

Table 10 presents the performance of the descriptors on the DUD data sets. Consistent with the previous results, the BEDROC scores on DUD data sets are much higher than in the cancer data sets. Further, due to the high BEDROC scores, there is not much scope for improvement. As can be seen, JPS-d performs as well or better on 15 out of the 20 data sets. JPS-d, along with Daylight and Molprint2D, perform consistently better than Joelib and 3pp. The consistent high BEDROC scores achieved by JPS-d, Molprint2D, and Daylight on all the DUD data sets suggest that the early part of the ranked list contains ligands that are easy to classify.

Overall, the series of experiments performed to benchmark the proposed descriptor show that JPS-d is highly efficient in molecular classification. Further, a key distinction between JPS-d and other descriptors is the set of the features employed to characterize a molecule. Fingerprinting techniques such Molprint2D, Daylight, or 3pp are all exhaustive in nature. On the other hand, JPS-d employs features that are statistically significant and thus can be used for other querying and mining tasks such as analyzing molecules based on how well they conform to a given pharmacophore model, similarity in the binding site of targets, correlating actives from cell-based assays to potential binding targets, and visualization of statistically significant geometrical patterns.

Correlation between Descriptors. To gain an intuition on the similarity of the rankings produced by each of the descriptors, we compute the Spearman's rank correlation coefficient (ρ). ρ is a measure of the statistical dependence between two ranked lists. It accesses how well the relationship between two ranked lists can be described using a monotonic function. Mathematically, the correlation between two ranked lists $X = [x_1,\cdot\cdot\cdot,x_n]$ and $Y = [y_1,\cdot\cdot\cdot,y_n]$ can be computed as follows

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad (7)$$

Table 11 presents the Spearman's rank correlation coefficients among JPS-d, Daylight, Molprint2D, 3pp, and Joelib on three of the cancer data sets. ρ is bounded within the range $[-1, 1]$, where -1 signifies a perfect negative correlation, 0 signifies no correlation, and 1 signifies a perfect positive correlation. As can be seen, JPS-d has a high correlation with 3pp since both descriptors are based on the same primitive features. A relatively high correlation also exists between Daylight and Molprint2D. The high correlation is likely due to both descriptors being based on 2D topological features.

The above results clearly establish the practical usefulness of mining the joint pharmacophore space for molecular analysis. The proposed clustering procedure and the statistical model successfully identify subspaces in the pharmacophore space that are discriminative in nature. The experiments showed that typically each significant subspace corresponds to a specific target. The discriminative potential of these subspaces is further verified by using them as keys of a molecular descriptor. These subspaces can be studied to gain a better understanding of the correlation between geometric features and biological activity and extend the state of the art in virtual screening technologies.

DISCUSSION

In this paper, we studied the problem of pharmacophore modeling from a new geometric perspective. We formulated the idea of a joint pharmacophore space and focused on developing efficient techniques to mine this space. Toward that goal, we derived efficient statistical models and a *k*-medoid clustering algorithm based on distances between triangles.

The examination of a joint space of ligands and proteins has been studied before. Current research has examined the joint binding space of ligands and proteins while measuring similarity of ligands through fingerprint-like or physicochemical descriptors and measuring the similarity of proteins through sequence similarity. Recently proposed methods and software tools have also approached the problem of 3D extraction and optimization of pharmacophores against a specific target.[57] Our proposed approach is unique in the definition of the joint pharmacophore space through the geometric arrangement of pharmacophoric features (donors, acceptors, hydrophobic cores, etc.) and the subsequent mining of this space to understand diversity, binding affinities, and biological properties. The novelty of geometric features allows us to consider different conformations of compounds and isolate the geometry of important pharmacophoric points in these conformations. The analysis of the space is now carried out through the spatial arrangement of these features, aspects that relate strongly to biological activity. The benefits of such joint pharmacophore analysis are numerous including increased sensitivity and specificity of results, flexibility in posing queries, scaffold hopping and early *in silico* prediction and pruning of compounds with undesirable properties.

In the empirical evaluation of the proposed methods, we considered only one conformation per molecule. It is natural to extend the setup to multiple conformations. While the underlying model to compute the significance of a cluster does not change, a number of interesting questions arise. Given conformations of a molecule and their activities (either known or predicted by a classifier), how should a class label be assigned to the molecule? Given the ranked list of conformations generated by a classifier (SVM in our case), how should one generate the ranked list of molecules? Currently, we make the assumption that as long as there is at least one active conformation of a molecule the molecule can be termed as active. Moreover, in the ranking procedure, the emphasis is on finding the best conformation of a molecule rather than aggregating all of its active conformations. To illustrate further, since classification is performed on conformations of molecules, it is necessary to develop a method that generates a ranked list of molecules from the ranked list of conformations generated by the classifier. In our approach, only the highest scoring conformation of a molecule is considered in computing its final rank. Certainly, alternative ranking mechanisms based on metrics such as average score of conformations or statistical significance of the conformation scores of a molecule can be pursued. In our future work, we plan to analyze such approaches.

At the core of our work lies the characterization of molecules using triangles. In the field of molecular classification, this approach has been studied in 3-point pharmacophores. However, there are a few fundamental differences in the transformation from the feature space to the vector representation of a molecule. First, 3pp is exhaustive in nature, and thus all triangles used as features may not be statistically interesting. On the other hand, we take a more selective approach by incorporating a sophisticated preprocessing step. Only triangles that are discriminative are selected as features for the molecular descriptor. The second key difference between 3pp and JPS-d stems from the density estimation capabilities of binning and clustering. Clustering is better able to identify dense and sparse regions in the pharmacophoric space and accordingly selects triangles to represent those regions. However, binning is blind to the density of objects in the space and thus may suffer from dividing a single cluster into multiple bins or enclosing multiple clusters in a single bin. Due to these key differences, triangles employed by JPS-d have higher information content. Further, the significant triangles mined from the joint pharmacophore space are capable of more than just molecular classification. As evident from the empirical evaluation of the clustering patterns, significant triangles indicate the geometric configurations of pharmacophoric features that correlate with a biological activity and thus can be used in various higher-order tasks such as categorization of molecules, annotating subspaces with biological tags, and top-*k* similarity searches. Molecular classification is just one of those various applications that can be developed on the platform provided by significant triangles.

Extensions of 3pp exist that are based on 4-point pharmacophores rather than three. It is thus natural to analyze such extensions of our proposed approach. A key weakness of larger pharmacophoric features stems from the combinatorial explosion in the size of the pharmacophore space and consequently reduced scalability. In our future work, we plan to explore the option of employing larger pharmacophoric features and yet maintain a pharmacophoric space that is more tractable. Such a technique can be designed, if a more selective procedure is employed in selecting the larger pharmacophoric features. More specifically, one can adopt an hierarchical approach where 4-point pharmacophoric features are generated by "joining" two 3-point pharmacophores that are statistically significant and tend to co-occur in the same conformation. In a similar manner, one can extend the technique to 5-point pharmacophores by joining statistically significant and co-occurring 4-point pharmacophores. Such a join-based approach of generating pharmacophoric features promises to create a pharmacophoric space that is smaller in size and denser in content.

Overall, the proposed approach is a notable deviation from existing pharmacophore modeling techniques. Applications of mining the joint pharmacophore space is evident from the patterns observed in the clustering results and its subsequent use in developing a molecular descriptor for classification. The novelty of geometric pharmacophoric features allows us to consider different conformations of compounds and isolate the geometry of important pharmacophoric points in these conformations. Moreover, the proposed descriptor based on significant triangles achieved better results than well-known techniques such as Daylight fingerprints and 3pp. The proposed formulation of the joint pharmacophore space opens up a new direction in the field of drug discovery by developing an improved beginning investigation point for medicinal chemists.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sayan@cs.ucsb.edu (S.R.); ambuj@cs.ucsb.edu (A.K.S.).

■ ACKNOWLEDGMENT

We thank Gregory D. Cuny for sharing the CDK-5 data set and helping us with some insightful comments during the preparation of the manuscript. The work was supported by NSF Grant IIS-0917149.

■ REFERENCES

(1) Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discovery Today* **2005**, *10*, 139–147.

(2) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2006**, *3*, 413–421.

(3) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382–384.

(4) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, 232–240.

(5) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

(6) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.

(7) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts smallmolecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(8) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(9) Wale, N.; Karypis, G. Target Fishing for Chemical Compounds Using Target-Ligand Activity Data and Ranking Based Methods. *J. Chem. Inf. Model.* **2009**, *49*, 2190–2201.

(10) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.

(11) Nidhi,; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.

(12) Niwa, T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J. Med. Chem.* **2004**, *47*, 2645–2650.

(13) Warmuth, M. K.; Liao, J.; RÃd'tsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2003.

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

(15) Jacob, L.; Vert, J. P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.

(16) Bock, J. R.; Gough, D. A. Virtual Screen for Ligands of Orphan G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402–1414.

(17) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt):A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

(18) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.

(19) Geppert, H.; Humrich, J.; Stumpfe, D.; Gartner, T.; Bajorath, J. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–779.

(20) Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions. *Bioinformatics* **2005**, *21*, 4289–4296.

(21) Lindstrom, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. Hierarchical PLS Modeling for Predicting the Binding of a Comprehensive Set of Structurally Diverse Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2006**, *46*, 1154–1167.

(22) Ning, X.; Rangwala, H.; Karypis, G. Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets. *J. Chem. Inf. Model.* **2009**, *49*, 2444–2456.

(23) Strombergsson, H.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Wikberg, J. E. S.; Kleywegt, G. J.; Hvidsten, T. R. Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space. *J. Chem. Inf. Model.* **2008**, *48*, 2278–2288.

(24) Weill, N.; Rognan, D. Development and Validation of a Novel Protein-Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049–1062.

(25) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. P. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.

(26) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des.* **2006**, *67*, 370–372.

(27) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(28) Kurogi, Y.; Guner, O. F. Pharmacophore modeling and three-dimensional database searching for drug design using catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035–1055.

(29) Guner, O.; Clement, O.; Kurogi, Y. Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. *Curr. Med. Chem.* **2004**, *11*, 2991–3005.

(30) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.

(31) Molecular Operating Environment (MOE). http://www.chemcomp.com/index.htm (accessed April 9, 2011).

(32) Diller, D. J. In Silico hERG Modeling: Challenges and Progress. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 106–121.

(33) *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Aliso Viejo, CA, 2008.

(34) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 1122–1133.

(35) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.

(36) Beno, B. R.; Mason, J. S. The design of combinatorial libraries using properties and 3D pharmacophore fingerprints. *Drug Discovery Today* **2001**, *6*, 251–258.

(37) van Dongen, S. Ph.D. Thesis, University of Utrecht, 2000.

(38) Kabsch, W. A solution of the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, 922.

(39) *Wolfram MathWorld*. http://mathworld.wolfram.com/BinomialDistribution.html (accessed November 16, 2010).

(40) Vapnik, V. *Statistical Learning Theory*; John Wiley: New York, NY, USA.

(41) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction ofmutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21*, 359–368.

(42) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug. Discovery* **2002**, *1*, 882–894.

(43) Bohm, H.-J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; John Wiley & Sons, Inc.: New York, NY, USA, 2000.

(44) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840–1848.

(45) Ranu, S.; Singh, A. K. Mining Statistically Significant Molecular Substructures for Efficient Molecular Classification. *J. Chem. Inf. Model.* **2009**, *49*, 2537–2550.

(46) *Pubchem*. (http://pubchem.ncbi.nlm.nih.gov (accessed November 16, 2010).

(47) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(48) *JOELib: a java based computational chemistry package*. http://joelib.sourceforge.net/ (accessed November 16, 2010).

(49) Ketchen, D. J.; Shook, C. L. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management J.* **1996**, *17*, 441–458.

(50) Yan, M. Ph.D. Thesis, Virginia Polytechnic Institute and State University, 2005.

(51) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.

(52) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2009. http://www.csie.ntu.edu.tw/cjlin/libsvm (accessed April 29, 2009).

(53) Truchon, J.-F. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model* **2007**, *47*, 488–508.

(54) Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* **2010**, *26*, 1348–1356.

(55) Zhao, W.; Hevener, K.; White, S.; Lee, R.; Boyett, J. A statistical framework to evaluate virtual screening. *BMC Bioinformatics* **2009**, *10*, 225.

(56) Clark, R.; Webster-Clark, D. Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.

(57) Langer, T. *Pharmacophores and Pharmacophore Searches*; John-Wiley and Sons: Chichester, 2006.