# Latent Semantic Structure Indexing (LaSSI) for Defining Chemical Similarity

Richard D. Hull,† Suresh B. Singh, Robert B. Nachbar, Robert P. Sheridan, Simon K. Kearsley, and Eugene M. Fluder*

*Department of Molecular Systems, RY50S-100, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065*

*Received September 6, 2000*

A novel method for computing chemical similarity from chemical substructure descriptors is described. This new method, called LaSSI, uses the singular value decomposition (SVD) of a chemical descriptor-molecule matrix to create a low-dimensional representation of the original descriptor space. Ranking molecules by similarity to a probe molecule in the reduced-dimensional space has several advantages over analogous ranking in the original descriptor space: matching latent structures is more robust than matching discrete descriptors, choosing the number of singular values provides a rational way to vary the "fuzziness" of the search, and the reduction in the dimensionality of the chemical space increases searching speed. LaSSI also allows the calculation of the similarity between two descriptors and between a descriptor and a molecule.

## Introduction

Pharmaceutical companies produce and license large numbers of chemical compounds. This valuable resource is even more valuable if it can be effectively mined for new leads. One mining method, similarity searching, starts with a "probe", a molecule with interesting biological activity. The goal is to find other molecules in chemical databases similar in structure to the probe, hopefully with similar activities. This idea was reified as the "similar property principle" by Johnson and Maggiora[1] in 1990, and similarity searching has become a standard tool for molecular modeling.[1−11] Such search methods rank database molecules by decreasing similarity to the probe. After computing this ranking, the user of the system typically selects some number of the top-ranked molecules for further study. This task is often complicated by the fact that the goal is not only to identify untested molecules that have the desired biological activity but also to find molecules that are structurally novel relative to the probe.

There are many possible representations of chemical structures, and the choice of representation is at least as important as the choice of similarity measure. Suitable representations include ones based upon topological (2D), geometrical (3D), and/or physicochemical (log *P*, surface area, etc.) descriptors.[4−7] Advantages and disadvantages are found for each of these descriptors. For example, while methods using topological representations are usually computationally less expensive than those using geometrical representations, topological descriptors may not be expressive enough to capture the stereochemistry or conformational properties of molecules. Recent efforts have investigated techniques to combine the results of several descriptor types.[4,8]

Much research has been carried out to explore new representation schemes and to compare their success in ranking active molecules from a database by their similarity to a probe or by their success in clustering actives.[9] One very practical approach to describing molecules is the vector space model popularized by Willett.[2] This method involves representing a molecule as a set of 2D or 3D substructures and their frequencies. Sometimes only the presence or absence of a substructure is noted, as in molecular fingerprint methods. Similarity searches with the vector representation are very practical because comparing lists of precomputed vectors is computationally inexpensive.

One property of vector models is that the substructure features are treated as completely independent. The presence of related but not identical descriptors does not make any contribution to increasing the similarity score between two molecules. This is contrary to the experience of medicinal chemists who readily recognize that certain features may be at least partially equivalent. Therefore, it is possible that treating descriptors as independent will cause some interesting molecules to be missed in similarity searches. The method we present in this manuscript, **La**tent **S**emantic **S**tructure **I**ndexing, or LaSSI, attempts to overcome this deficiency by automatically uncovering related descriptors and using them in the calculation of similarity.

In this paper we present the following: (1) the mathematical underpinnings of the LaSSI approach, which were inspired by latent semantic indexing, a method used in document retrieval; (2) the methodology for calculating chemical similarity; (3) extensions to the basic methodology; (4) a straightforward small example to illustrate LaSSI. A detailed analysis of the application

* To whom correspondence should be addressed. Tel: 732-594-5074. Fax: 732-594-4224. E-mail: fluder@merck.com.
† Present address: Elagent Corp., 7011 N. Atlantic Ave., Suite 200, Cape Canaveral, FL 32920.

10.1021/jm000393c CCC: $20.00 
Published on Web 04/05/2001

of LaSSI to searching a large database of drug-like molecules can be found in our companion paper.[12]

## Mathematical Background

The mathematical underpinnings of LaSSI were inspired by **L**atent **S**emantic **I**ndexing (LSI), a document retrieval technique originally described in Deerwester et al.[13] LSI addresses a problem plaguing keyword search algorithms, that of synonymy. Synonymy is the phenomenon that many words in English, and in other languages, have similar meanings. This becomes a problem when the user of such a system chooses a different synonym in his/her query than was used by the authors of relevant documents. For example, a search for documents about automobiles using the keyword "automobile" would miss documents that do not use that term but instead use the terms "car", "motor-car", "motor-vehicle", or "horseless carriage". The problem is that the string of characters "car" does not match the string of characters of the term "automobile"; hence, car and automobile are treated as unrelated terms. Synonymy is typically handled through the construction of thesauri that are later used to expand the query to include the synonyms of each term. Unfortunately, thesauri are difficult to build, and one is never sure they are complete or contain no errors.

LSI alleviates this problem by automatically uncovering statistical relationships between the terms found in a collection of documents. For instance, "automobile" and "car" would each co-occur in the same documents with words such as "gasoline", "road", and "engine." This would point to a relationship of "automobile" and "car". These relationships can then be used to calculate similarities between terms and documents and between documents and documents that exploit related terms as well as exact matches. It is our belief that this approach can be used to overcome the problem of related descriptors in the chemical domain. We will now describe the mathematics behind LSI and LaSSI.

LSI represents a collection of text documents as a term−document matrix. LaSSI, on the other hand, uses a chemical descriptor−molecule matrix. Hence, the nature of the input matrices for LaSSI and LSI are very different, but the mathematical treatment of these matrices is the same. Later we will see that the calculation of similarities made by LSI and LaSSI is related but somewhat different.

A collection of molecules in a chemical database is initially represented as a set of vectors, where each vector $\mathbf{v}_i = (d_{1i}, d_{2i}, ..., d_{ni})^T$ consists of the nonnegative frequency of occurrence of each descriptor $j$ in molecule $i$ and where $n$ is the total number of uniquely occurring descriptors in the entire set of molecules. A chemical descriptor−molecule matrix, **X**, therefore, is a set of two or more such vectors, i.e., $\mathbf{X} = \{\mathbf{v}_1, ..., \mathbf{v}_m\}$, where the number of molecules $m \geq 2$, or

$$\mathbf{X} = \begin{matrix} \\ \text{descriptors} \end{matrix} \overset{\text{molecules}}{\begin{bmatrix} d_{11} & d_{21} & \cdots & d_{m1} \\ d_{12} & d_{22} & \cdots & d_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{mn} \end{bmatrix}}$$

Two implicit choices have already been made that we will now attempt to justify. First, the choice to use the raw frequency of descriptor occurrence has been challenged in the document retrieval literature. Modifications of the frequency counts by multiplying them by the inverse document frequency (IDF), calculated as the log(collection size/number of documents containing the term) or various measures of informational entropy, have been reported to improve the recall and precision of ranked documents using these techniques.[14] It is our belief that these modifications were motivated by the poor performance of some queries for which we have a different remedy and which will be subsequently explained. Furthermore, we have not experienced the same improvements in recall and precision in the chemical similarity domain and hence are satisfied with the raw descriptor frequencies.

Second, no pretreatment of the matrix **X**, such as mean centering, variance scaling, or vector size normalization, is performed. Each value of the matrix represents an integer descriptor frequency and thus the scales of the descriptors are commensurate; scaling and centering are unnecessary. Also, size normalization, calculated by dividing each $d_{ji}$ by the total length of the molecule $i$, is not performed because our goal is often to find novel molecules whose size differs from the original probe molecule(s).

LaSSI employs the singular value decomposition (SVD) of **X** to produce a reduced-dimensional representation of the original matrix. The SVD technique is well-known in the linear algebra literature[15] and has been used in many scientific[16,17] and engineering[18] applications, including signal and spectral analysis. Here we show a novel application of SVD to the problem of calculating chemical similarity, though a report describing the use of SVD to support the visualization of chemical similarity has been recently reported.[19]

Let the SVD of **X** in $R^{m \times n}$ be defined as $\mathbf{X} = \mathbf{P}\Sigma\mathbf{Q}^T$ where **P** is a $n \times r$ matrix, called the left singular matrix ($r$ is the rank of **X**), and its columns are the eigenvectors of $\mathbf{XX}^T$ corresponding to non-zero eigenvalues. **Q** is a $m \times r$ matrix, called the right singular matrix, whose columns are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to non-zero eigenvalues. $\Sigma$ is a $r \times r$ diagonal matrix = diag($\sigma_1$, $\sigma_2$, ..., $\sigma_r$) whose non-zero elements, called singular values, are the square roots of the eigenvalues and have the property that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$:

$$\underset{\mathbf{X}}{\begin{bmatrix} d_{11} & d_{21} & \cdots & d_{m1} \\ d_{12} & d_{22} & \cdots & d_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ d_{1n} & d_{2n} & \cdots & d_{mn} \end{bmatrix}} \xrightarrow{\text{SVD}} \underset{\mathbf{P}}{\begin{bmatrix} p_{11} & \cdots & p_{r1} \\ p_{12} & \cdots & p_{r2} \\ \vdots & \vdots & \vdots \\ p_{1n} & \cdots & p_{rn} \end{bmatrix}} \bullet \underset{\Sigma}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}} \bullet \underset{\mathbf{Q}^T}{\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ q_{r1} & q_{r2} & \cdots & q_{rm} \end{bmatrix}}$$

The $k$th rank approximation of $\mathbf{X}$, $\mathbf{X}_k$, for $k < r$, where $\sigma_{k+1} \ldots \sigma_r$ are set to 0, can be efficiently computed using variants of the Lanczos algorithm.[20] $\mathbf{X}_k$, called the partial SVD of $\mathbf{X}$, is the matrix of rank $k$ which is the closest to $\mathbf{X}$ in the least-squares sense. It is generated by:

$$\mathbf{X}_k = \mathbf{P}_k \Sigma_k \mathbf{Q}_k{}^T$$

The rows of $\mathbf{X}_k$ are orthogonal "latent" descriptors that are linear combinations of the original descriptors, and the columns are the projection of the molecules into the space of those descriptors. In practice, however, similarities between molecules and/or descriptors are more easily calculated from $\mathbf{P}_k$ and/or $\mathbf{Q}_k$, and the actual construction of $\mathbf{X}_k$ is unnecessary (see Methods).

Deerwester et al.[13] showed that given the partial SVD of $\mathbf{X}$ in LSI, it is possible to compute similarities between terms, between documents, and between a term and a document. Furthermore, they could compute the similarity of a new document (a column vector that does not exist in $\mathbf{X}$) to both the terms and the documents in the database. In terms of chemistry, a new molecule not already in the database, say a probe, can be added by first projecting it into the $k$-dimensional space of the partial SVD. The projection of a probe vector $\mathbf{z}$ would be $\mathbf{y} = \mathbf{z}^T \mathbf{P}_k \Sigma_k{}^{-1}$. $\mathbf{y}$ can then be treated as a row of $\mathbf{Q}_k$ for the purposes of calculating similarity (see below).

Chemistry-flavored LSI calculations of similarity would be as follows: The similarity of two descriptors $i$ and $j$ is calculated by computing the dot product between the $i$th and $j$th rows of the matrix $\mathbf{P}_k \Sigma_k$. The similarity of two molecules $i$ and $j$ can be calculated by computing the dot product between the $i$th and $j$th rows of the matrix $\mathbf{Q}_k \Sigma_k$. The similarity of a descriptor $i$ to a molecule $j$ can be calculated by computing the dot product between the $i$th row of the matrix $\mathbf{P}_k \Sigma_k$ and the $j$th row of the matrix $\mathbf{Q}_k \Sigma_k$. The use of $\Sigma_k$ means that the dimensions are scaled by the singular values. In contrast, LaSSI does not use scaling. Therefore, the calculation of LaSSI similarity between two entities is as follows:

$$\text{LaSSI similarity between descriptors } i \text{ and } j = \sum_x p_{ix} p_{jx} / |\mathbf{p}_i||\mathbf{p}_j|$$

$$\text{LaSSI similarity between descriptor } i \text{ and molecule j} = \sum_x p_{ix} q_{jx} / |\mathbf{p}_i||\mathbf{q}_j|$$

$$\text{LaSSI similarity between molecules } i \text{ and } j = \sum_x q_{ix} q_{jx} / |\mathbf{q}_i||\mathbf{q}_j|$$

where $x$ goes from 1 to k. $p_{ix}$ and $q_{ix}$ are elements of $\mathbf{P}_k$ and $\mathbf{Q}_k$, respectively. $\mathbf{p}_i$ is a row of $\mathbf{P}_k$ and $q_{ix}$ is an element of $\mathbf{Q}_k$.

This type of similarity, the normalized dot product of two vectors, is often called the cosine similarity because the similarity index (−1 to 1) is equal to the cosine of the angle of the vectors formed by the entities (molecules or descriptors) relative to the origin of the space. Ignoring the scaling component $\Sigma_k$ in LaSSI improves the system's ability to select similar molecules regardless of whether the probe's descriptors are well represented in the database. We will discuss this advantage in greater detail in the section describing method extensions.

## Methods

There are two distinct phases of processing: (1) constructing a LaSSI version of a chemical database and (2) calculating the similarity of the molecules of the LaSSI database to the probe molecule(s). The first phase is computationally expensive. However, it only needs to be performed once to create the database. The second phase, on the other hand, can be accomplished very quickly − a search of an average-sized database ($\sim 10^5$ molecules) can be performed in under 1 min on a modest computer workstation. This section describes the details of both phases.

**Constructing a LaSSI Database.** Constructing a LaSSI database requires compiling chemical descriptors for each molecule represented in the database, creating an index relating the columns of the matrix to the molecules and another index relating the rows of the matrix to the chemical descriptors, creating a chemical descriptor−molecule matrix representing the molecules in the chemical database, and performing the SVD of this matrix.

The creation of a descriptor−molecule matrix is straightforward. First one must decide on what descriptors to use. In our experience, the topological descriptors atom pairs and topological torsions have worked extremely well. Atom pairs (APs)[10] are substructures of the form:

$$\text{AT}_i - \text{AT}_j - (\text{distance})$$

where (distance) is the distance in bonds along the shortest path between an atom of type $\text{AT}_i$ and an atom of type $\text{AT}_j$. Atom types encode the species of atom, the number of bonded non-hydrogen neighbors, and the number of $\pi$ electrons. For instance the descriptor type "n21o1005" would mean a nitrogen with 2 non-hydrogen neighbors and one $\pi$ electron five bonds away from an oxygen with one neighbor and no $\pi$ electrons. Topological torsions (TTs)[11] are of the form:

$$\text{AT}_i - \text{AT}_j - \text{AT}_k - \text{AT}_l$$

where $i$, $j$, $k$, and $l$ are consecutively bonded distinct atoms and the atom types are as described above. All of the APs and/or TTs in a molecule are counted to form a frequency vector. We have also experimented with vectors of 3D geometric descriptors, combinations of 2D and 3D descriptors, and biological descriptors such as $IC_{50}$'s for specific receptors. However, for the purposes of this paper we will restrict our discussion to AP and TT descriptors.

An in-house topological descriptor generator[4] is used to generate AP and TT descriptors from the connection table of each molecule in the chemical database. A first pass through the database is performed to create a catalog of unique descriptors and another catalog of each molecule name. Then, a second pass creates a list of the frequency of each descriptor found in each molecule. Recall that the value of matrix element $d_{ji}$ of $\mathbf{X}$ is the frequency of descriptor $j$ in molecule $i$. The resulting matrix is used as input for public domain SVD routines[20] which produce the partial SVD of the matrix. We generally keep the 1000 largest singular values and vectors for a LaSSI database. The database consists of the singular values and right and left singular vectors produced by the SVD.

**Querying a LaSSI Database.** Querying a LaSSI database is carried out as follows: A user specifies a probe molecule. The connection table of the probe molecule (or multiple molecules in the case of a joint probe) is converted into the descriptor set of the LaSSI database to create a column vector $\mathbf{z}$ for the probe. This vector is then projected into the reduced $k$-dimensional space as $\mathbf{y}$, as described in the Mathematical

Background section, for some $k$ specified by the user. The normalized dot products of each molecule vector with the transformed probe **y** are calculated and the resulting values are sorted in descending order, maintaining the index of the molecule responsible for that value. The user is then presented with a list of the top-ranked molecules truncated at a user-defined number, e.g., usually the highest-ranked 300, 500, or 1000 molecules.

By varying the number of singular values (the choice of $k$), the user can control the level of fuzziness of the search: larger values of $k$ produce better approximations of the original descriptor space than smaller values. In the limiting case of $k = r$, $r$ being the rank of the matrix $\mathbf{X}$, $\mathbf{X}_k = \mathbf{X}$ and all the descriptors are independent. Alternatively, if $k$ is small, much of the discrete character of the descriptors will be lost and similarity will be more "fuzzy". As we will see, the decision of what value of $k$ to use depends on the task at hand. A later section describes includes a discussion of a particular kind of tuning called singular value calibration wherein the best value of $k$ is found.

**Joint Probes.** LaSSI can easily handle "joint probes", that is probes that are the descriptor average of two or more molecules. For each descriptor $j$,

$$\mathbf{z}_j = \frac{\sum_{i=1}^{M} d_{ji}}{M}$$

where $M$ is the number of molecules in the probe. In LaSSI, the cosine similarity measure is identical whether or not the sum is divided by $M$, so the joint probe is equivalent to the strategy of summing frequencies from the document retrieval literature. Relevance feedback is a technique of expanding queries for retrieving text documents from a database that involves summing the terms found in successfully retrieved documents with the terms in the original query.[14] Studies have shown significant increases in the precision and recall of document retrieval when relevance feedback is used.[21] An in-depth analysis of the use of joint probes for selecting actives from a large drug-like chemical database is presented in another paper in this series,[22] and that study shows analogous improvements.

**Singular Vector Calibration.** One question that must be answered, whether the probe consists of a single molecule or several molecules, is how many singular vectors, $k$, to use in the similarity calculation. Published reports of the use of LSI for searching document databases have typically used 100−300 singular vectors for modest to large databases.[13,23] Instruction in the selection of $k$ has been limited in this area to the following rule of thumb: "choose small values of $k$ for conceptual or fuzzy searches, larger values for more literal searches". Investigations of LSI have shown poor performance for some queries when an arbitrary $k$ had been chosen.

Our approach has been to calibrate the selection of $k$ to the probe at hand whenever more than one active is known for a given biological activity. Similarity calculations for $k = 10$, 20, 30, ..., $k_{max}$ are performed and the rank of each known active is combined to create a composite score called the "initial enhancement".[4] Initial enhancement is defined as how many more actives are found in the top $N$-ranked molecules than would be expected by chance given the total number of actives. The value of $k$ with the highest initial enhancement is retained as $k_{best}$. Presumably this is the value of $k$ that optimizes searching for new actives. Typically $k_{max} = 1000$ for a large chemical database. We usually use $N = 300$. This is a somewhat arbitrary but reasonable number; 300 molecules can be easily tested in a moderate throughput biological assay. In our companion paper[12] we show that rankings with $k_{best}$ are 20−30% better than when using a static default value of $k$.

One might wonder why initial enhancements should be sensitive to $k$ and why $k_{best}$ depends on the probe and the biological activity of interest. The first reason is that the descriptors in only a fraction of the singular vectors are relevant to the biological activity. The remaining singular vectors can be seen as adding noise. The consequence of this is that an arbitrarily chosen $k$ might work for some probes and not for others. The effect of calibration is more dramatic for LaSSI than for LSI. In LSI the singular values $\Sigma_k$ are used to scale the singular vectors, so that the descriptors in all but the first few eigenvectors make a negligible contribution. In contrast, LaSSI does not scale the singular vectors. This means that if the 500th singular vector captures an important relationship between a probe and other molecules in the database with a similar biological activity, it will not be deprecated to near obscurity by the difference in size of the 500th singular value and the 499 preceding ones. Moreover, users of LaSSI often present probes that contain descriptors not found with great frequency in the database. By not scaling the singular vectors, it is possible to use rare descriptors when that is necessary.

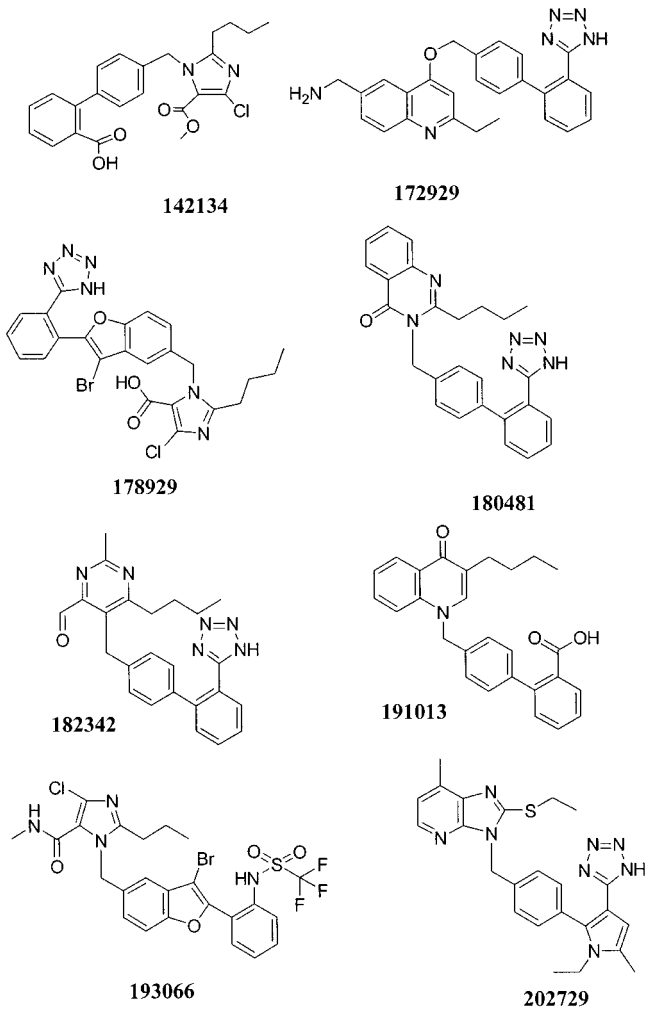


**Figure 1.** Eight angiotensin II antagonists from the MDL Drug Data Report database.

**Descriptor Coloring.** Descriptor coloring is a means of understanding which parts of highly ranked molecules are responsible for that ranking. A feature of using SVD is that the similarity between molecules and descriptors can be calculated directly. Each descriptor has a similarity to the probe, and these similarities can be used to color the atoms of ranked molecules. The contribution of each descriptor is calculated from the products of the SVD − we assign an atomic coefficient to atom $i$ as $a_i = \sum_j \mathrm{sim}_j$, where $j$ runs over the descriptors (APs and/or TTs) that contain atom $i$. For LaSSI $\mathrm{sim}_j$ is the cosine similarity of descriptor $j$ to the probe. These atomic coefficients are then used to highlight atoms, either by color or by radius.

**Table 1.** Portion of the Descriptor−Molecule Matrix for Eight Angiotensin II Antagonists (the complete matrix has dimensions 286 × 8)

| | 142134 | 172929 | 178929 | 180481 | 182342 | 191013 | 193066 | 202729 |
|---|---|---|---|---|---|---|---|---|
| c10c1007 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| c20br1005 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| c20br1008 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| c20br1009 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| c20c1001 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| c20c1002 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| c20c1003 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| c20c1005 | 1 | 0 | 0 | 0 | 2 | 0 | 2 | 2 |
| c20c1006 | 2 | 1 | 1 | 1 | 2 | 0 | 1 | 1 |
| c20c1007 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| c20c1008 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| c20c2001 | 2 | 0 | 2 | 2 | 2 | 2 | 1 | 0 |
| c20c2002 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| c20c2003 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| c20c2004 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| c20c2005 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| c20c2006 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| c20c2007 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| c21br1003 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| c21br1004 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 |
| c21br1005 | 0 | 0 | 3 | 0 | 0 | 0 | 3 | 0 |
| c21br1006 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| c21c1003 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| c21c1004 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| c21c1005 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 4 |
| c21c1006 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 4 |
| c21c1007 | 2 | 0 | 0 | 1 | 3 | 1 | 4 | 6 |
| c21c1008 | 4 | 2 | 2 | 4 | 4 | 2 | 2 | 4 |
| c21c1009 | 2 | 2 | 1 | 3 | 2 | 3 | 0 | 0 |
| c21c1011 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| c21c1012 | 3 | 1 | 0 | 1 | 3 | 0 | 2 | 4 |
| c21c1013 | 3 | 2 | 1 | 2 | 3 | 1 | 4 | 2 |
| c21c1014 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 0 |
| c21c1015 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| c21c2002 | 2 | 5 | 2 | 2 | 2 | 4 | 2 | 2 |
| c21c2003 | 2 | 4 | 1 | 2 | 3 | 4 | 1 | 3 |
| c21c2004 | 0 | 2 | 0 | 2 | 1 | 4 | 0 | 3 |
| c21c2005 | 2 | 4 | 2 | 7 | 3 | 4 | 2 | 3 |
| c21c2006 | 5 | 2 | 3 | 10 | 6 | 7 | 4 | 3 |
| c21c2007 | 6 | 4 | 4 | 9 | 6 | 9 | 2 | 3 |
| c21c2008 | 3 | 5 | 3 | 4 | 3 | 6 | 2 | 0 |
| c21c2009 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 0 |
| c21c2010 | 3 | 0 | 1 | 3 | 3 | 1 | 1 | 0 |
| c21c2011 | 4 | 1 | 3 | 4 | 4 | 3 | 3 | 2 |
| c21c2012 | 3 | 3 | 4 | 3 | 3 | 4 | 3 | 1 |

**Table 2.** Portion of the (286 × 8) **P** Matrix

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0055 | 0.0222 | 0.0001 | −0.0297 | −0.0068 | −0.0246 | 0.0210 | −0.0321 |
| 2 | 0.0053 | −0.0039 | 0.0343 | 0.0189 | 0.0009 | 0.0016 | −0.0217 | −0.0164 |
| 3 | 0.0053 | −0.0039 | 0.0343 | 0.0189 | 0.0009 | 0.0016 | −0.0217 | −0.0164 |
| 4 | 0.0053 | −0.0039 | 0.0343 | 0.0189 | 0.0009 | 0.0016 | −0.0217 | −0.0164 |
| 5 | 0.0251 | 0.0107 | 0.0115 | −0.0290 | 0.0287 | −0.0071 | 0.0266 | 0.0035 |
| 6 | 0.0167 | −0.0195 | 0.0145 | −0.0100 | −0.0431 | 0.0167 | −0.0234 | −0.0124 |
| 7 | 0.0169 | −0.0016 | 0.0021 | −0.0285 | −0.0263 | 0.0360 | 0.0517 | −0.0130 |
| 8 | 0.0180 | 0.0245 | 0.0375 | −0.0789 | −0.0050 | −0.0891 | −0.0419 | −0.0124 |
| 9 | 0.0247 | 0.0158 | 0.0159 | −0.0269 | −0.0562 | −0.0391 | −0.0008 | 0.0865 |
| 10 | 0.0076 | −0.0335 | 0.0112 | −0.0210 | 0.0061 | −0.0233 | −0.0176 | 0.0156 |
| 11 | 0.0106 | 0.0202 | 0.0008 | −0.0337 | 0.0580 | −0.0287 | 0.0669 | 0.0780 |
| 12 | 0.0311 | −0.0340 | 0.0122 | −0.0171 | −0.0973 | 0.0509 | 0.0036 | −0.0196 |
| 13 | 0.0143 | −0.0145 | −0.0023 | −0.0073 | −0.0542 | 0.0342 | 0.0270 | −0.0072 |
| 14 | 0.0138 | −0.0009 | 0.0237 | −0.0059 | −0.0516 | 0.0177 | −0.0393 | 0.0290 |
| 15 | 0.0193 | −0.0066 | 0.0189 | −0.0310 | −0.0151 | 0.0185 | 0.0013 | −0.0182 |
| 16 | 0.0175 | −0.0101 | −0.0140 | 0.0159 | −0.0383 | 0.0069 | 0.0276 | 0.0203 |
| 17 | 0.0061 | −0.0142 | −0.0210 | 0.0192 | 0.0245 | −0.0283 | 0.0165 | −0.0139 |
| 18 | 0.0057 | 0.0173 | −0.0074 | 0.0020 | 0.0439 | −0.0256 | 0.0253 | 0.0217 |
| 19 | 0.0053 | −0.0039 | 0.0343 | 0.0189 | 0.0009 | 0.0016 | −0.0217 | −0.0164 |
| 20 | 0.0106 | −0.0077 | 0.0685 | 0.0377 | 0.0018 | 0.0033 | −0.0435 | −0.0328 |
| 285 | 0.0053 | −0.0039 | 0.0343 | 0.0189 | 0.0009 | 0.0016 | −0.0217 | −0.0164 |
| 286 | 0.0082 | −0.0027 | 0.0518 | 0.0402 | −0.0094 | 0.0208 | 0.0069 | −0.0277 |

**Table 3.** Complete (8 × 8) Σ Matrix

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 124.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 36.37 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 32.21 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 23.88 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 19.08 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 15.11 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 13.53 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11.71 |

**Table 4.** Complete (8 × 8) $\mathbf{Q}^T$ Matrix

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.2806 | 0.3944 | 0.3617 | 0.4177 | 0.3608 | 0.3701 | 0.2982 | 0.3227 |
| −0.3615 | 0.1608 | 0.0417 | 0.1292 | 0.3394 | −0.6761 | −0.1842 | 0.4685 |
| 0.1213 | −0.3791 | 0.5643 | −0.3235 | −0.1371 | −0.2983 | 0.5391 | 0.1414 |
| −0.3482 | 0.5550 | 0.5093 | −0.0392 | −0.2027 | −0.0945 | −0.0590 | −0.5062 |
| −0.2616 | 0.3043 | −0.1958 | −0.0787 | −0.6623 | 0.1640 | 0.2133 | 0.5323 |
| −0.0734 | −0.4130 | 0.2892 | 0.7147 | −0.3986 | −0.0150 | −0.2643 | 0.0265 |
| 0.2281 | 0.0086 | 0.3875 | −0.4160 | −0.0493 | 0.2149 | −0.6817 | 0.3337 |
| 0.7271 | 0.3218 | −0.1319 | 0.1125 | −0.3078 | −0.4845 | −0.0601 | −0.0681 |

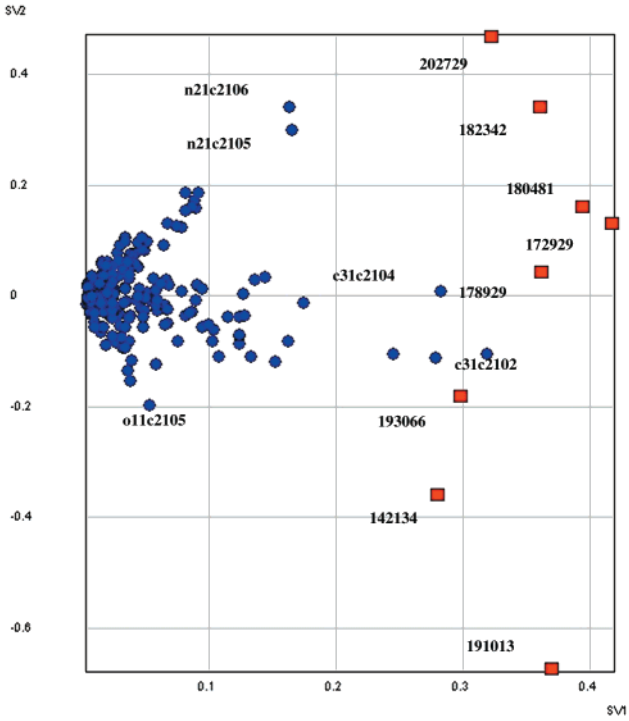


**Figure 2.** Plot of molecules and descriptors in two dimensions. Molecules are represented by red squares, descriptors by blue circles.

## Results

The concepts described previously can be best understood in the context of a small, illustrative example. Angiotensin II plays a critical role in regulation of fluid and electrolyte balance and arterial blood pressure. Increased activity of the renin−angiotensin system can result in hypertension and disorders of fluid and electrolyte homeostasis.[24] Two subtypes of the angiotensin II receptor, $AT_1$ and $AT_2$, located in the plasma membrane of cells are found in varying proportions at several sites within the body including vascular smooth muscle, adrenal cortex, brain, and kidney. Nonpeptide angiotensin II antagonists which have activity against $AT_1$ have been used to treat hypertension in human beings. Eight such molecules described in the MDL Drug Data Report (MDDR) database[25] are shown in Figure 1. Atom pair descriptors were calculated for each of these molecules. Each descriptor found in at least two different molecules was kept to create the descriptor−molecule matrix, of which a portion is shown in Table 1. Applying the SVD to this matrix results in the three matrices shown in Tables 2−4. By setting $k = 2$, we can plot the values of the first two singular vectors for each molecule and descriptor in two dimensions. Figure 2 shows a plot of descriptors and molecules. It is interesting to note that the most nitrogen-rich molecule 202729 is nearly collinear with the origin and the descriptors n21c2106 and n21c2105 (using atoms from the tetrazole and the biphenyl), while the molecule 191013 is collinear with the descriptor o11c2105. Descriptors n21c2106 and n21c2105 are just one pair of descriptors that are highly correlated in this set of molecules.

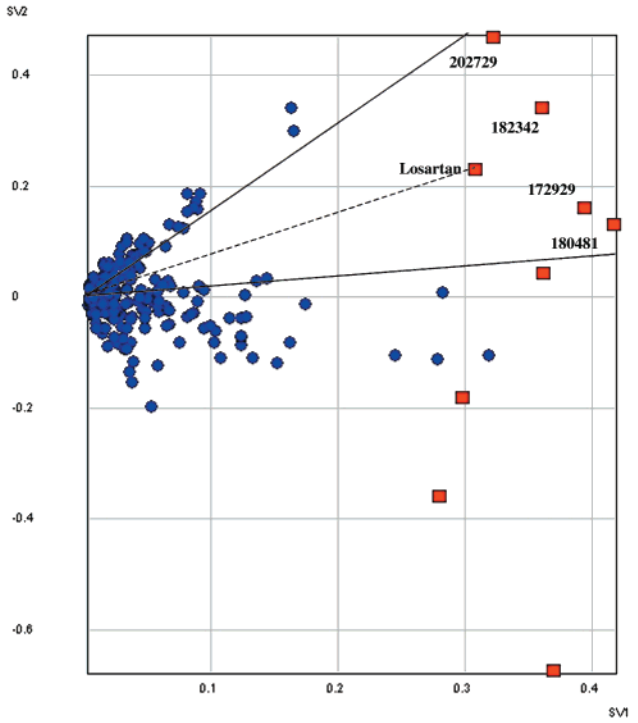


**Figure 3.** Probe molecule losartan is projected into two dimensions, and molecules with a cosine similarity of >0.9 are labeled.

We can calculate the pairwise similarity matrix for the eight molecules using $\mathbf{Q}_k$. We can also project a ninth molecule, different from the original eight, into the 2D space and then calculate the similarity of the eight molecules to this probe, losartan (structure in Figure 4). Descriptors unique to losartan are ignored. The result of the projection of losartan into the space of the other eight molecules is shown in Figure 3. Cosine similarity between each of the eight A-II antagonists and losartan can be computed, producing the following ranking:

| rank | molecule | cosine similarity |
|---|---|---|
| 1 | 182342 | 0.994 |
| 2 | 172929 | 0.968 |
| 3 | 202729 | 0.947 |
| 4 | 180481 | 0.942 |
| 5 | 178929 | 0.865 |
| 6 | 193066 | 0.372 |
| 7 | 142134 | 0.019 |
| 8 | 191013 | −0.140 |

The molecules similar to losartan at >0.9 are emphasized in Figure 3.

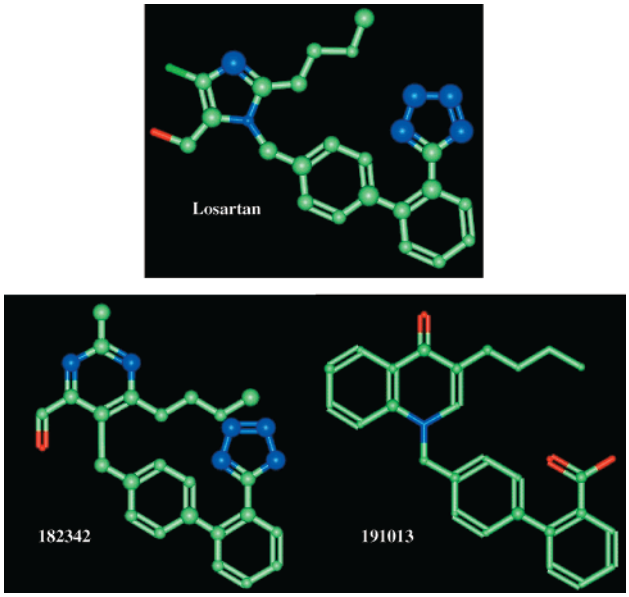


**Figure 4.** Descriptor colorings of the most and least similar molecules (184342, 191013) to losartan. Radius of the atoms indicates similarity between the descriptors containing those atoms and the losartan probe.

One of the issues raised in the Introduction was that chemical descriptors are correlated and that some descriptors count more than others in the reduced-dimensional space. This is illustrated in Figure 4, wherein "coloration" is done against losartan as the probe at $k = 2$. The radii of the atoms are proportional to $a_i$. First we note than when losartan is compared against itself as a probe, not all atoms are equally emphasized. In particular, the tetrazole nitrogens and the $sp^2$ nitrogen in the imidazole appear to count more. We can examine the relationships LaSSI has uncovered with the other molecules by looking at the top-ranked molecule (182342) and the last-ranked molecule (191013) from the table above. Losartan and 182342 are very similar, but again not all parts of 182342 are equally emphasized. The tetrazole and pyrimidine nitrogens appear to count more than any of the carbon atoms in the biphenyl. On the other hand, we see that nearly all the atoms of 191013 have a small radius, consistent with that molecule having a small similarity to the probe.

## Discussion

We have extended a method (LSI), which was developed for calculating the similarity between text documents, to create an analogous method (LaSSI) that can be applied to molecules. This approach inherently assumes that the chemical descriptors in a database are not independent but associated, much as terms in a document are associated. Using the singular value decomposition of an appropriately constructed descriptor−molecule matrix, one can uncover these associations and use them to calculate chemical similarity. In cases where more than one compound of interest is known, joint probes and singular value calibration can be used to improve LaSSI's performance. Finally, the user can color the atoms of probe and database molecules to reveal those structural components responsible for a molecule's similarity score.

LSI and LaSSI as originally formulated are designed to start with integer frequencies, of terms and substructure descriptors, respectively. However, any type of real number descriptors can be used in addition to or in place of integer frequencies. The only additional complication with real number descriptors is that, not being naturally commensurate with each other, they usually have to be normalized before processing.

There are certain similarities of LaSSI to principal components analysis[26] wherein molecules are projected into reduced-dimensional space defined by orthogonal axes. Indeed SVD can be used to calculate principal components.[19] Sometimes principal components analysis is used to visualize chemical data and is then restricted to low numbers of dimensions. In other applications, molecules are clustered in the low-dimensional space. The principal component axes and the number of principal components depend only on the variance in the structural data of the molecules. Activity data is generally, although not always,[27] ignored. When that is the case, the most relevant descriptors to activity may not be in the principal components that are retained. In LaSSI, we retain many of the singular vectors and can adjust the number of vectors to get the best correlation with similarity in that space with activity. While principal components analysis lets us calculate the similarity between molecules, LaSSI allows us also to calculate the similarity between descriptors and between descriptors and molecules.

The example we show here involves only a few molecules, but LaSSI can be routinely applied to tens or hundreds of thousands of molecules. In our companion paper[12] we show the application of LaSSI to searching a large database of drug-like moleclues.

**Acknowledgment.** The authors thank J. Chris Culberson, Kate Holloway, and Ralph Mosley for providing in-house examples for testing LaSSI. Phil Harris worked on theoretical comparison of SVD with PCA. Mike Berry, at the University of Tennessee, Knoxville, provided the implementation of SVD that we currently use.

## References

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley: New York, 1990.
(2) Willett, P. *Similarity and clustering in chemical information systems*; Research Studies Press Ltd., John Wiley & Sons: New York, 1987; 254 pp.
(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(4) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(5) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparisons of fragment-based measures of shape similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(6) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(7) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1−66.
(8) Ginn, C. M. R.; Ranade, S. S.; Willett, P.; Bradshaw, J. The Application of Data Fusion to Similarity Searching in Chemical Databases. Proceedings of the International Conference On Multisource-Multisensor Information Fusion, Fusion'98, 1998, pp 307−313.
(9) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Clustering Methods and Descriptors for use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(10) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-acitivity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(11) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(12) Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. P.; Kearsley, S. K.; Sheridan, R. P. Chemical Similarity Searches using Latent Semantic Structure Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **2001**, *44*, 1185−1191.
(13) Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landuaer, T. K.; Harshman R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391−407.
(14) Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, 1983.
(15) Strang, G. *Linear Algebra and its Applications*, 3rd ed.; Harcourt Brace & Co.: New York, 1988.
(16) Tsurui, H.; Nishimura, H.; Hattori, S.; Hirose, S.; Okumura, K.; Shirai, T. Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition. *J. Histochem. Cytochem.* **2000**, *48*, 653−662.
(17) Kojima, M.; Tanokura, M.; Maeda, M.; Kimura, K.; Amemiya, Y.; Kihara, H.; Takahashi, K. pH-dependent unfolding of aspergillopepsin II studied by small-angle X-ray scattering. *Biochemistry* **2000**, *39*, 1364−1372.
(18) Paul, J. S.; Reddy, M. R.; Kumar, V. J. A transform domain SVD filter for suppression of muscle noise artifacts in exercise ECG's. *I. E. E. E. Trans. Biomed. Eng.* **2000**, *47*, 654−663.
(19) Xie, D.; Tropsha, A.; Schlick, T. An Efficient Projection Protocol for Chemical Databases: the Singular Value Decomposition Combined with Truncated-Newton Minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 167−177.
(20) Berry, M.; Do, T.; O′Brien, G.; Krishna, V.; Varadhan, S. *SVDPAKC Version 1.0 User Guide UTK Technical Report CS-93-194*; Computer Science Department, University of Tennessee: Knoxville, TN, April 1993; revised March 1996.
(21) Salton, G.; Buckley, C. Improving retrieval performance by relevance feedback. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 288−297.
(22) Singh, S. B.; Sheridan, R. P.; Fluder, E. M.; Hull, R. D. Mining the Chemical Quarry with Joint Chemical Probes: An application of Latent Semantic Structure Indexing (LaSSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem.*, in press.
(23) Dumais, S. Latent Semantic Indexing (LSI): TREC-3 Report. TREC-3 Proceedings, Gaithersburg, MD, 1994; pp 219−230.
(24) Smith, R. D.; Chiu, A. T.; Wong, P. C.; Herblin, W. F.; Timmermans, P. B. Pharmacology of nonpeptide angiotensin II receptor antagonists. *Annu. Rev. Pharmacol. Toxicol.* **1992**, *32*, 135−165.
(25) MDDR-3D, version 98.1, is available from MDL Information Systems, Inc., San Leandro, CA.
(26) Glen, W.; Dunn, W. J., III; Scott, D. R. Principal components analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(27) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.

JM000393C