Data Visualization during the Early Stages of Drug Discovery

Dharmesh M. Maniyar,*,† Ian T. Nabney,† Bruce S. Williams,‡ and Andreas Sewing‡

Neural Computing Research Group, Information Engineering, Aston University, Birmingham, B4 7ET, United Kingdom, and Lead Discovery Technologies, Pfizer Global Research, Sandwich, CT13 9NJ, United Kingdom

Received October 27, 2005

Multidimensional compound optimization is a new paradigm in the drug discovery process, yielding efficiencies during early stages and reducing attrition in the later stages of drug development. The success of this strategy relies heavily on understanding this multidimensional data and extracting useful information from it. This paper demonstrates how principled visualization algorithms can be used to understand and explore a large data set created in the early stages of drug discovery. The experiments presented are performed on a real-world data set comprising biological activity data and some whole-molecular physicochemical properties. Data visualization is a popular way of presenting complex data in a simpler form. We have applied powerful principled visualization methods, such as generative topographic mapping (GTM) and hierarchical GTM (HGTM), to help the domain experts (screening scientists, chemists, biologists, etc.) understand and draw meaningful decisions. We also benchmark these principled methods against relatively better known visualization approaches, principal component analysis (PCA), Sammon's mapping, and self-organizing maps (SOMs), to demonstrate their enhanced power to help the user visualize the large multidimensional data sets one has to deal with during the early stages of the drug discovery process. The results reported clearly show that the GTM and HGTM algorithms allow the user to cluster active compounds for different targets and understand them better than the benchmarks. An interactive software tool supporting these visualization algorithms was provided to the domain experts. The tool facilitates the domain experts by exploration of the projection obtained from the visualization algorithms providing facilities such as parallel coordinate plots, magnification factors, directional curvatures, and integration with industry standard software.

I. INTRODUCTION

High-throughput screening (HTS) is an important part of early stages of the drug discovery process for many pharmaceutical companies.[1] Despite its wide use, high attrition rates in the later stages of drug discovery have raised questions about the viability of the high-throughput paradigm.[2,3] It is being recognized that increasing the quality of screening libraries and HTS assays, rather than their quantity, is likely to be an important determinant for the identification of active compounds that have a chance to make it through the drug discovery pipeline.[2,4] The choice of active compounds from biological screening is no longer focused solely on potency but relies on data from multiple sources (physicochemical properties, selectivity data, etc.), which are gathered in parallel to the target-directed measurements.[5] Clustering, visualization, and other computational methods can be useful to select and understand biologically active compounds according to their receptor affinity during the early stages of drug discovery.[6−8]

There are many virtual screening techniques widely applied in the drug discovery process,[9−13] for example, for the selection of compounds for HTS,[14] the design of target-family focused combinatorial libraries,[15] or the improvement of adverse physicochemical or toxicity issues in the lead optimization phase.[16,17] Extensive efforts have been made to develop and apply computational methods for the classification of active molecules.[18−20] To enhance speed, reduce costs, and provide insight during the discovery of novel drugs, it is crucial to find the molecular descriptors within a set of active molecules which are responsible for the affinity and selectivity toward a specific biological target or a target family. Thus, it is important to understand the results from the past HTS screens on that particular target or target family and relate the results to other properties of molecules.

Data visualization is an important means of extracting useful information from large quantities of raw data. It is difficult for people to visualize data in more than three dimensions, so high-dimensional data is projected onto lower-dimensional space. Here, we use the term *visualization* to mean any method of projecting data into a lower-dimensional space in such a way that the projected data keeps most of the topographic properties (i.e., "structure") and makes it easier for the users to interpret the data to gain useful information from it. In this paper, we apply new principled data visualization techniques to facilitate the domain experts' understanding of past screening results and we also discuss how the results from these algorithms could help to prioritize compounds from the virtual compound library for future assay development.

Traditional visualization techniques such as principal component analysis (PCA),[21] Sammon's mapping,[22] and self-organizing maps (SOMs)[23] have been applied in the drug discovery community for visualization and dimensionality reduction.[24−27] As demonstrated in this paper, simple linear projection by PCA or an attempt to preserve relative similarities between the higher-dimensional input data and projected lower-dimensional data by NeuroScale[28] (a neural network implementation of Sammon's mapping) are not

* Corresponding author phone: +44 784 356 7510; fax: +44 121 204 3685; e-mail: maniyard@aston.ac.uk.
† Aston University.
‡ Pfizer Global Research.

10.1021/ci050471a CCC: $33.50
Published on Web 05/26/2006

adequate for good clustering of the vast multidimensional data one has to deal with during the early stages of the drug discovery process. Even the SOM, which is a neural network architecture for unsupervised learning, did not prove effective. Instead, generative topographic mapping (GTM),[29] which is a principled probabilistic nonlinear projection, yielded better clustering for the problem at hand.

Loosely speaking, GTM identifies the computer screen with a two-dimensional "rubber sheet" that is injected into the high-dimensional data space. The sheet is supposed to "cover" the cloud of data points by locally stretching, contracting, and curving. The visualization plot is obtained by first projecting the data points onto the rubber sheet and then letting the rubber sheet relax to its original form of the computer screen. We refer to the injected (possibly curved and stretched) two-dimensional rubber sheet in the data space as the *projection manifold*. Besides the visualization plot itself, the user is often interested in additional information about the structure of the projection manifold in the high-dimensional data space. It is possible to calculate local magnification factors[30] for the GTM which describe how small regions on the computer screen are stretched or compressed when mapped to the data space. Magnification factors can be used for detecting (on the visualization plot) separate clusters in the data space. The GTM's *smooth* two-dimensional projection manifold allows us to analytically describe (local) geometric properties anywhere on the manifold. It is not possible to obtain magnification factors for PCA and Sammon's mapping. For the SOM, the magnification factors can only be approximated.[31] When injecting a two-dimensional rubber sheet into a high-dimensional data space, the projection manifold may form complicated folds that cannot be detected by using magnification factors alone. It is also possible in GTM to analytically calculate the local directional curvatures of the projection manifold to provide the user with a facility for monitoring the amount of folding and neighborhood preservation in the projection manifold.[32]

Moreover, it has been argued that a single two-dimensional projection, even if it is nonlinear, is not usually sufficient to capture all of the interesting aspects of a large high-dimensional data set. Hierarchical extensions of visualization methods allow the user to "drill down" into the data; each plot covers a smaller region, and it is therefore easier to discern the structure of the data. The amount of data one has to deal with during the early stages of drug discovery is vast and high-dimensional. HGTM is a hierarchical GTM visualization system which allows the user to explore interesting regions in more detail.[33]

We have provided a software tool supporting advanced information visualization facilities to assist the domain experts to interactively explore the projection obtained with the visualization algorithm. The domain experts can study the structure and properties of interesting molecules on the projection manifold using this tool. The tool can be used in combination with other widely used software tools in the pharmaceutical domain such as Pipeline Pilot[34] and SpotFire.[35]

In this paper, we demonstrate two applications of the principled visualization algorithms on a large high-dimensional data set: (1) understand and explore a large data set containing past HTS results (biological activity data) alongside some whole-molecular properties and (2) make decisions to prioritize compounds from the virtual compound library for future HTS campaigns. This uses data containing physicochemical properties only.

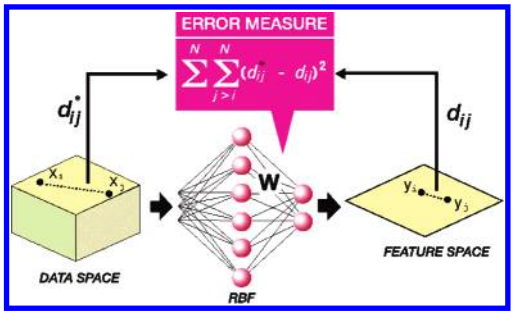

Figure 1. Schematic representation of the NeuroScale model.

The remainder of this paper is organized as follows: In section II, we provide an overview of the visualization techniques we used. The details of the interactive software we provide to the domain experts are given in section III. The description of the data sets we considered for this paper is given in section IV. Results are reported in section V, followed by a discussion in section VI. Finally, we conclude the paper with conclusions in section VII.

II. VISUALIZATION ALGORITHMS

We applied the following five visualization algorithms.

A. Principal Component Analysis (PCA). PCA is a multivariate procedure which rotates the data such that maximum variabilities are projected onto the axes. Essentially, a set of correlated variables is transformed into a set of uncorrelated variables which can be ordered by reducing variability. The uncorrelated variables are linear combinations of the original variables, and in many cases, the last few of these variables can be removed with a minimum loss of information.[21]

The first principal component is the combination of variables that explains the greatest amount of variation. Subject to being orthogonal, the second principal component defines the next largest amount of variation to the first principal component. If the covariance matrix has full rank, there can be as many principal components as there are variables.

Given a set of observations $\mathbf{x}^{(i)} \in \mathscr{R}^d$, $i = 1, ..., N$, which are centered, $\sum_{i=1}^{N} \mathbf{x}^{(i)} = 0$, in PCA, we find the principal component by diagonalizing the covariance matrix

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}\mathbf{x}^{(i)\mathrm{T}} \tag{1}$$

Because we want to project multidimensional data on two dimensions, we use the first two principal components. The main disadvantage is that the technique is linear; therefore, any nonlinear correlation between variables will not be captured.

B. NeuroScale. Neuroscale[28] is a novel neural network implementation of Sammon's mapping. It is a projection method which attempts to preserve the relative "similarity" (in this case, characterized by *Euclidean* distance) between points in the data space ($\mathbf{x}$) and latent space ($\mathbf{y}$), as shown in Figure 1 (adapted from ref 36).

The topographic nature of the transformation is imposed by the "STRESS" term, which accounts for the preservation of interpoint similarity

$$E = \sum_{i}^{N}\sum_{j>i}^{N}(d_{ij}^{*} - d_{ij})^2 \tag{2}$$

$$d_{ij}^{*} = ||\mathbf{x}_i - \mathbf{x}_j|| \tag{3}$$

$$d_{ij} = ||\mathbf{y}_i - \mathbf{y}_j|| \tag{4}$$

Points are projected from the data space onto the latent space by means of a radial basis function (RBF) neural network which adapts to the required projection mapping. The RBF parameters are defined by minimizing the "STRESS" term, rather than the traditional residual error minimization. where $\mathbf{W}$ is a weight matrix and $\phi(\ )$ are the basis functions.

$$\mathbf{y} = \mathbf{W}\,\phi(\mathbf{x}) \tag{5}$$

C. Self-Organizing Map (SOM). The SOM is a popular unsupervised learning algorithm, based on a grid of artificial neurons whose weights are adapted to match input vectors in a training set.[23] There are many variants of SOM training algorithms. The batch version of the SOM algorithm can be described as follows.

A set of reference vectors $\mathbf{z}_i$ is defined in the data space, in which each vector is associated with a node on a regular lattice in a (typically) two-dimensional "feature map". The algorithm begins by initializing the reference vectors (for example, by setting them to random values, by setting them equal to a random subset of the data points, or by using principal component analysis). Each cycle of the algorithm then proceeds as follows. For every data vector $\mathbf{t}_n$, the corresponding "winning node" $j(n)$ is identified, corresponding to the reference vector $\mathbf{z}_j$ having the smallest Euclidean distance $||\mathbf{z}_j - \mathbf{t}_n||^2$ to $\mathbf{t}_n$. The reference vectors are then updated by setting them equal to weighted averages of the data points given by

$$\mathbf{z}_i = \frac{\sum_n h_{ij(n)}\,\mathbf{t}_n}{\sum_n h_{ij(n)}} \tag{6}$$

in which h_{ij} is a neighborhood function associated with the ith node. This is generally chosen to be a unimodal function of the feature map coordinates centered on the winning node, for example, a Gaussian. The steps of identifying the winning nodes and updating the reference vectors are repeated iteratively. A key ingredient in the algorithm is that the width of the neighborhood function h_{ij} starts with a relatively large value and is gradually reduced after each iteration.

D. Generative Topographic Mapping (GTM). GTM models a probability distribution in the (observable) high-dimensional data space, $\mathscr{D} = \mathscr{R}^D$, by means of low-dimensional latent, or hidden, variables.[29] The data are visualized in the latent space, $\mathscr{H} \subset \mathscr{R}^L$.

As demonstrated in Figure 2 (adapted from ref 29), we cover the latent space, $\mathscr{H}$, with an array of K latent space centers, $\mathbf{x}_i \in \mathscr{H}$, $i = 1, 2, ..., K$. The nonlinear GTM transformation, f: $\mathscr{H} \Rightarrow \mathscr{D}$, from the latent space to the data

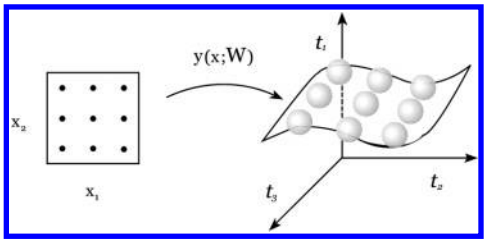

Figure 2. Schematic representation of the GTM model.

space is defined using a RBF network. To this end, we cover the latent space with a set of $M - 1$ fixed nonlinear basis functions (we use Gaussian functions of the same width σ), ϕ: $\mathscr{H} \Rightarrow \mathscr{R}$, $j = 1, 2, ..., M - 1$, centered on a regular grid in the latent space. Given a point $\mathbf{x} \in \mathscr{H}$ in the latent space, its image under the map f is

$$f(\mathbf{x}) = \mathbf{W}\,\phi(\mathbf{x}) \tag{7}$$

where $\mathbf{W}$ is a $D \times M$ matrix of weight parameters and $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), ..., \phi_M(\mathbf{x})]^T$.

GTM creates a generative probabilistic model in the data space by placing a radially symmetric Gaussian with zero mean and inverse variance β around images, under f, of the latent space centers $\mathbf{x}_i \in \mathscr{H}$, $i = 1, 2, ..., K$. We refer to the Gaussian density associated with the center $\mathbf{x}_i$ by $P(\mathbf{t}|\mathbf{x}_i,\mathbf{W},\beta)$. Defining a uniform prior over $\mathbf{x}_i$, the density model in the data space provided by the GTM is $P(\mathbf{t}|\mathbf{W},\beta) = {}^1/_K\sum_{i=1}^{K} P(\mathbf{t}|\mathbf{x}_i,\mathbf{W},\beta)$.

For the purpose of data visualization, we use Bayes' theorem to invert the transformation f from the latent space $\mathscr{H}$ to the data space $\mathscr{D}$. The posterior distribution on $\mathscr{H}$, given a data point $\mathbf{t}_n \in \mathscr{D}$, is a sum of delta functions centered at centers $\mathbf{x}_i$, with coefficients equal to the posterior probability $R_{i,n}$ that the ith Gaussian (corresponding to the latent space center $\mathbf{x}_i$) generated $\mathbf{t}_n$[29]

$$R_{i,n} = \frac{P(\mathbf{t}_n|\mathbf{x}_i,\mathbf{W},\beta)}{\sum_{j=1}^{K} P(\mathbf{t}_n|\mathbf{x}_j,\mathbf{W},\beta)} \tag{8}$$

The latent space representation of the point $\mathbf{t}_n$, that is, *the projection of* $\mathbf{t}_n$, is taken to be the mean, $\sum_{i-1}^{K} R_{in}\mathbf{x}_i$, of the posterior distribution on $\mathscr{H}$.

The f image of the latent space $\mathscr{H}$, $\Omega = f(\mathscr{H}) = \{f(\mathbf{x}) \in \mathscr{R}^D\ |\mathbf{x} \in \mathscr{H}\}$, forms a smooth L-dimensional manifold in the data space. We refer to the manifold Ω as the *projection manifold* of the GTM.

Magnification factors of Ω are calculated as the Jacobian of the GTM map f in ref 30. Tiňo et. al.[32] derived a closed-form formula for directional curvatures of the projection manifold Ω for a latent space point $\mathbf{x} \in \mathscr{H}$ and a directional vector $\mathbf{h} \in \mathscr{H}$.

E. Hierarchical GTM (HGTM). The hierarchical GTM arranges a set of GTMs and their corresponding plots in a tree structure $\mathscr{T}$. An example HGTM structure is shown in the Figure 3. In this section, we give a general formulation of hierarchical mixture models; more detail can be found in ref 37.

The *Root* of the hierarchy is at level 1, that is, *Level(Root)* = 1. Children of a model $\mathscr{N}$ with $Level(\mathscr{N}) = \ell$ are at level $\ell + 1$, that is, $Level(\mathscr{M}) = \ell + 1$, for all $\mathscr{M} \in Children(\mathscr{N})$. Each model $\mathscr{M}$ in the hierarchy, except for *Root*, has an associated parent-conditional mixture coefficient, or prior π-

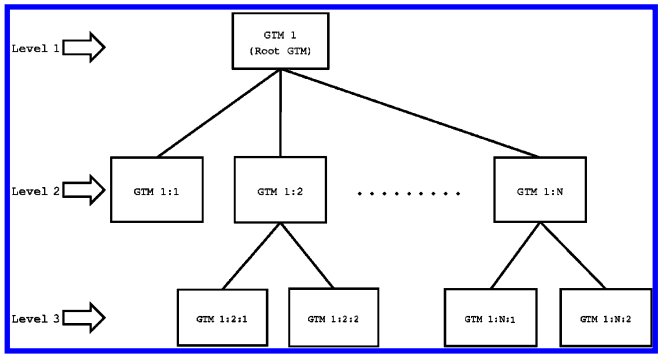

Figure 3. Example structure for HGTM.

($\mathcal{M}|Parent(\mathcal{M})$). The priors are non-negative and satisfy the consistency condition: $\sum_{\mathcal{M}\in Children(\mathcal{N})}\pi(\mathcal{M}|\mathcal{N}) = 1$. Unconditional priors for the models are recursively calculated as follows: $\pi(Root) = 1$, and for all other models

$$\pi(\mathcal{M}) = \prod_{i=2}^{Level(\mathcal{M})} \pi(Path(\mathcal{M})_i|Path(\mathcal{M})_{i-1}) \qquad (9)$$

where $Path(\mathcal{M}) = (Root, ..., \mathcal{M})$ is the N-tuple [$N = Level(\mathcal{M})$] of the nodes defining the path in $\mathcal{T}$ from *Root* to $\mathcal{M}$.

The distribution given by the hierarchical model is a mixture of leaf models of $\mathcal{T}$

$$P(t|\mathcal{T}) = \sum_{\mathcal{M}\in Leaves(\mathcal{T})} \pi(\mathcal{M})P(t|\mathcal{M}) \qquad (10)$$

Nonleaf models not only play a role in the process of creating the hierarchical model but in the context of data visualization can be useful for determining the relationship between related subplots in the hierarchy.

The hierarchical GTM is trained using the EM algorithm to maximize its likelihood with respect to the data sample $\zeta = \{t_1, t_2, ..., t_N\}$. Training of a hierarchy of GTMs proceeds in a recursive fashion. First, a base (*Root*) GTM is trained and used to visualize the data. Then, the user identifies interesting regions on the visualization plot that they would like to model in greater detail. In particular, the user chooses a collection of points, $c_i \in \mathcal{K}$, by clicking on the plot. These "regions of interest" given by these points (centers) are then transformed into the data space as Voronoi compartments[38] defined by the mapped points $f_{Root}(c_i) \in \mathcal{D}$, where f_{Root} is the map of the *Root* GTM. The child GTMs are initiated by local PCA in the corresponding Voronoi compartments. After training the child GTMs and seeing the lower-level visualization plots, the user may decide to proceed further and model in greater detail some portions of the lower-level plots and so forth. At each stage of the construction of a hierarchical GTM, the EM algorithm alternates between the E and M steps until convergence is satisfactory (typically, after 10−20 iterations).

III. THE SOFTWARE TOOL

An interactive software tool supporting these visualization algorithms is provided to the domain experts. The tool facilitates the domain experts to thoroughly explore the projection obtained from the visualization result providing facilities such as a parallel coordinates[39] technique, magnification factors,[30] and directional curvatures.[32] Using the parallel coordinates facility, a user can study molecules' properties in the data space while working with the projection obtained with the visualization algorithms. The parallel coordinates technique maps the D-dimensional data space onto the two display dimensions by using D equidistant axes which are parallel to one of the display axes. It displays each multidimensional data point as a polygonal line which intersects the horizontal dimension axes at the position corresponding to the data value for the corresponding dimension. Figure 4 displays how points in the visualization plot (highlighted in the plot) can be explored in the data space using the parallel coordinates technique.

One of the main features of GTM-based models is that it is possible to analytically calculate the magnification factors and the directional curvature of a GTM projection manifold. Magnification factor plots are used to observe the amount of stretching in different regions of the GTM projection manifold. A magnification factor plot also helps in outlier detection and cluster separation. Directional curvature plots allow the user to observe folding in the GTM projection manifold. The use of magnification factors and directional curvatures is demonstrated in section V.

The software also provides a way to integrate output of the visualization algorithms with industry standard software such as PartiView,[40] Pipeline Pilot, and SpotFire. Once the projection is obtained, the user can select data points from the projection manifold and study them further using the familiar advanced techniques available in these tools. The software is developed in MATLAB[41] using the NETLAB[42] toolbox. A stand-alone version of the software is available for Microsoft Windows and GNU/LINUX platforms.

IV. THE DATA

The data set provided by Pfizer is composed of "% response" data for five different targets obtained from different HTS screens, overall campaigning approximately 1 million compounds. A total of 44 862 compounds out of the total 1 million compounds were screened for all five targets. From this full-matrix data set of 44 862 compounds, 11 800 compounds were randomly selected to create the final data set (this reduced number of compounds was used for a clearer presentation of the results).

The five screens are of different target types. Two are peptidergic G-protein coupled receptor (GPCR) targets, two are aminergic GPCR targets, and the fifth screen is a kinase target. The four GPCR targets are related receptor types, while the kinase is a completely unrelated enzyme target class.

All results discussed in this paper have the activity threshold set at 40 [i.e., a compound is classified as active for a target if the screening result ("% response" value) for that target is greater than 40%, otherwise it is classified as an inactive for that target]. The threshold is only used for binning a compound as active or inactive to color it in the projection plots. The user can select any threshold value.

Table 1 lists the label (marker) information and distribution of compounds in different labels in our data set. Note that the markers specified in the Table are used in all of the projection plots presented in this paper.

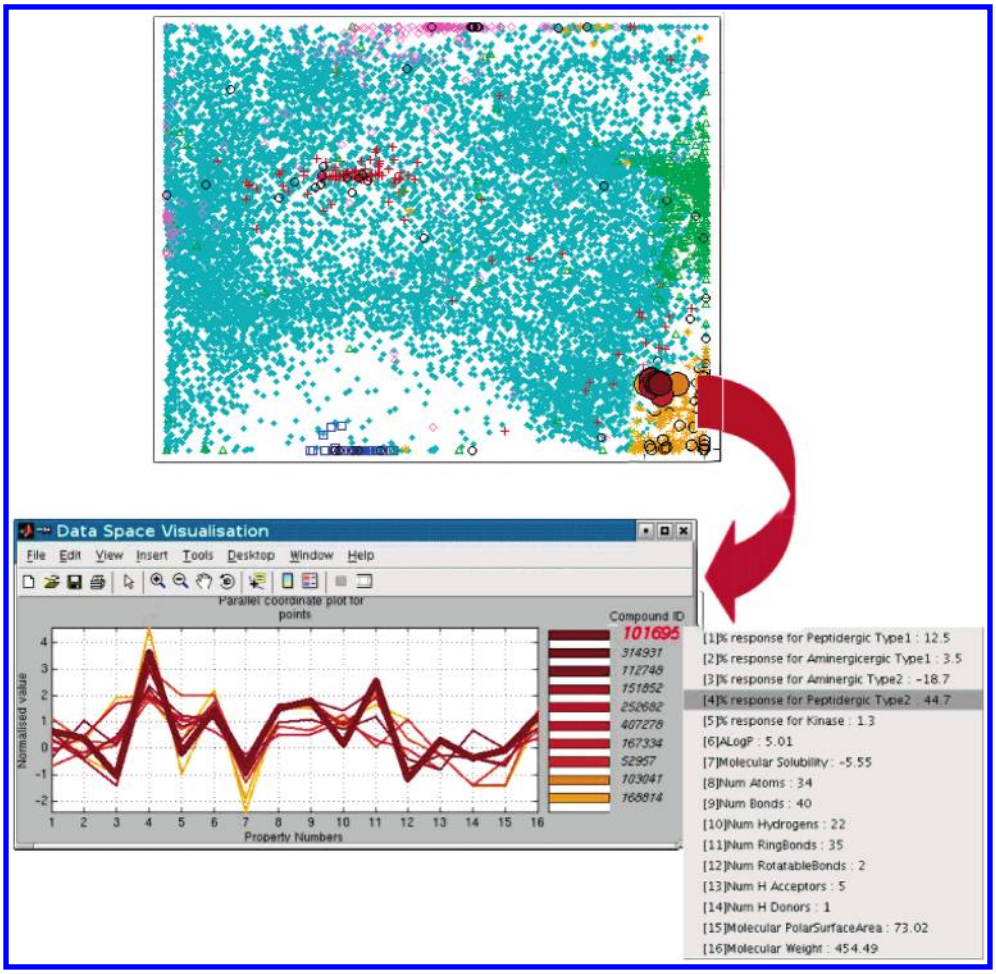

Figure 4. Example of parallel coordinates used on a GTM projection. When a user clicks on a point in the projection (upper plot), the data space visualization graph shows a color-coded plot of normalized property values for a group of compounds close in the projection space. The user can highlight a compound profile by selecting its compound ID on the right of the graph (see example ID highlighted in red). Actual data space values for the selected compound are displayed in a list at the right edge of the graph.

Table 1. Label Information and Compound Distribution Across Labels

Label Description	Marker	Compounds
Not active in any screen	•	10769
Active for peptidergic type1	+	118
Active for peptidergic type2	*	181
Active for aminergic type1	□	50
Active for aminergic type2	△	409
Active for kinase	◇	206
Active for more than 1 screen	○	66

Table 2. Molecular Physicochemical Properties

AlogP
molecular solubility
number of atoms
number of bonds
number of hydrogens
number of ring bonds
number of rotatable ring bonds
number of hydrogen acceptors
number of hydrogen donors
molecular polar surface area
molecular weight

In addition to the "% response" values obtained from the HTS campaign for these five targets, 11 whole-molecule physiochemical properties (listed in Table 2) are available for each compound in the data set.

Because different input variables in the data set have different ranges, before the development of visualization models, we apply a linear transformation (*Z*-score transformation) to have similar ranges for all of the variables. Each variable is treated independently and is rescaled as follows:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n \tag{11}$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_i^n - \mu_i)^2 \tag{12}$$

where $n = 1, \ldots, N$ indexes the patterns and μ_i and σ_i^2 represent the mean and variance of variable i respectively. Then, the values are scaled by

$$\tilde{x}_i^n = \frac{x_i^n - \mu_i}{\sigma_i} \tag{13}$$

where $\tilde{x}_i^n$ is the scaled value of variable i for the pattern n. Figure 5 shows histograms of all 16 variables after the scaling. These show approximately normal distributions.

V. RESULTS

A. Application 1. Using principled visualization algorithms, we can understand and explore a large data set containing past HTS results (biological activity data) alongside the

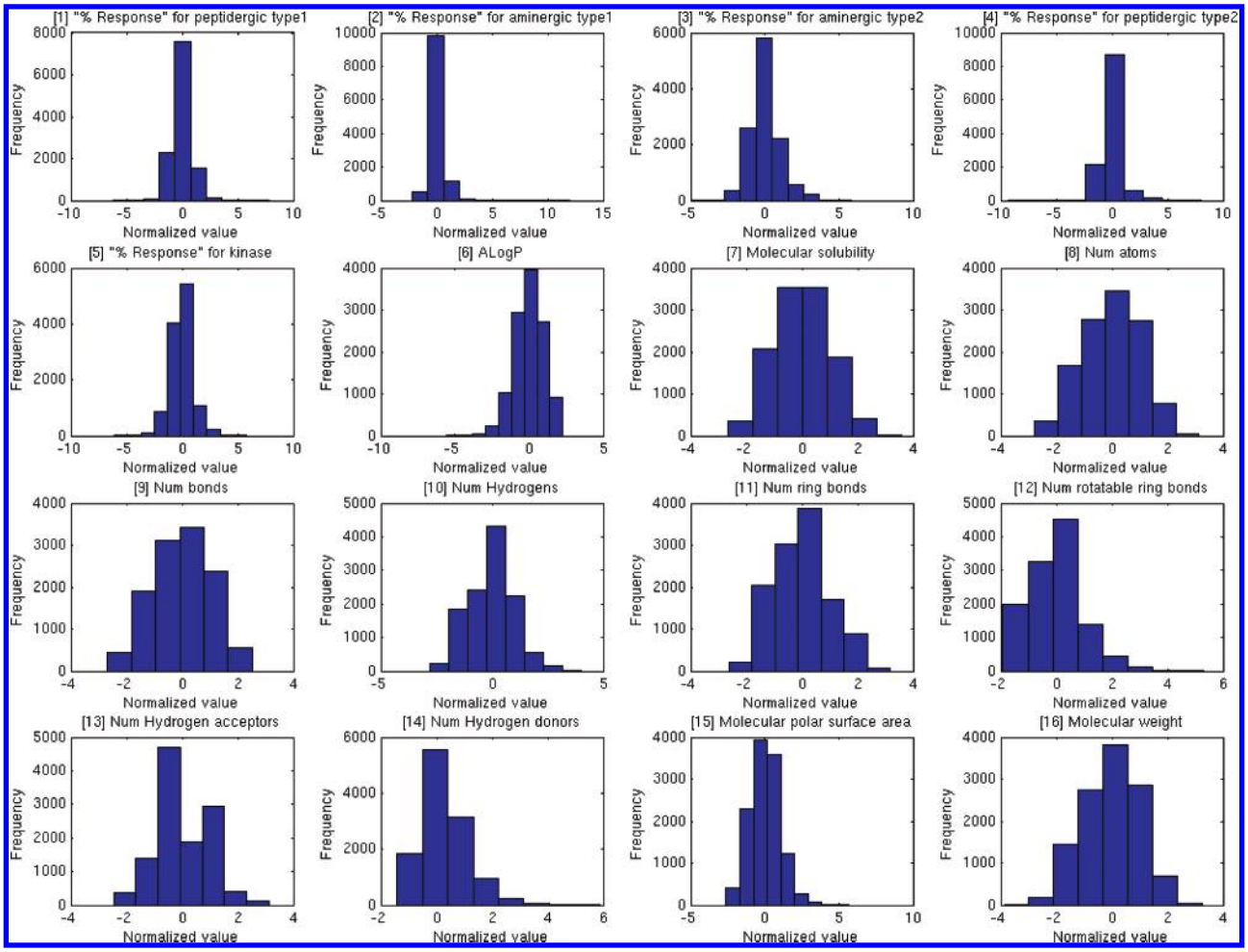

Figure 5. Histogram of scaled input variables.

whole-molecular properties. To demonstrate this, the experiments are carried out with the data set described in the previous section with all 16 input variables. In the rest of this section, we elaborate on the results we obtained using all of the visualization algorithms described in section II.

Not surprisingly, for a high-dimensional vast data set such as this, PCA, NeuroScale (a neural network implementation of Sammon's mapping), and SOM did not prove effective. The projections obtained using PCA (on the first two significant principal components), NeuroScale, and SOM visualization are presented in Figure 6a−c, respectively. The visualization results of all three of these algorithms do not show any useful separation of the active compounds from the inactive compounds. The projected data is like a "blob", and there is no apparent clustering; this does not give much insight into the detailed structure of the data. Active compounds are distributed all over the plots instead of forming any cluster, and thus, the results are not very useful to understand the data set.

The GTM visualization results are shown in Figure 6d. The GTM plot shows clear clusters for the compounds active for different targets and is certainly more informative. It is easier to understand and explore the data space using the GTM projection compared to the projection we obtained using traditional visualization techniques used in drug discovery.

Local parallel coordinate plots help us to observe variations in the patterns in different regions of a projection plot. Figure 7 shows how patterns of physicochemical properties vary in different regions of the GTM projection. A careful study with the parallel coordinate technique reveals interesting structures in the projection space. It can be observed that the active compounds for different targets are nicely clustered. The compounds active for peptidergic type 1 (marked as a red +) and peptidergic type 2 (marked as an orange *) targets are respectively clustered at the middle and bottom-right of the GTM projection plot (Figure 6d). Close study using the software tool reveals that the compounds marked as ○ present in the clusters for peptidergic type 1 and peptidergic type 2 are the active compounds for both of the peptidergic targets. That is in line with the fact that some compounds are active for both of the peptidergic targets. The compounds active for aminergic type 1 (marked as a purple □) and aminergic type 2 (marked as a green △) targets are respectively clustered separately at bottom-left and middle-right of the GTM projection plot (Figure 6d). The compounds active for the kinase target (marked as a pink ◇) are mostly clustered at the top-middle of the plot. Such different clusters are useful to understand the diversity of compounds for different targets. The compounds active for more than one target (marked as ○) are useful to observe overlaps and to understand the similarity of compounds active for different targets. It was observed that many inactive compounds (marked as a blue ●) near the active compounds have activity values near 40% (the threshold we set), which means they are close to being active. They are mostly the borderline cases

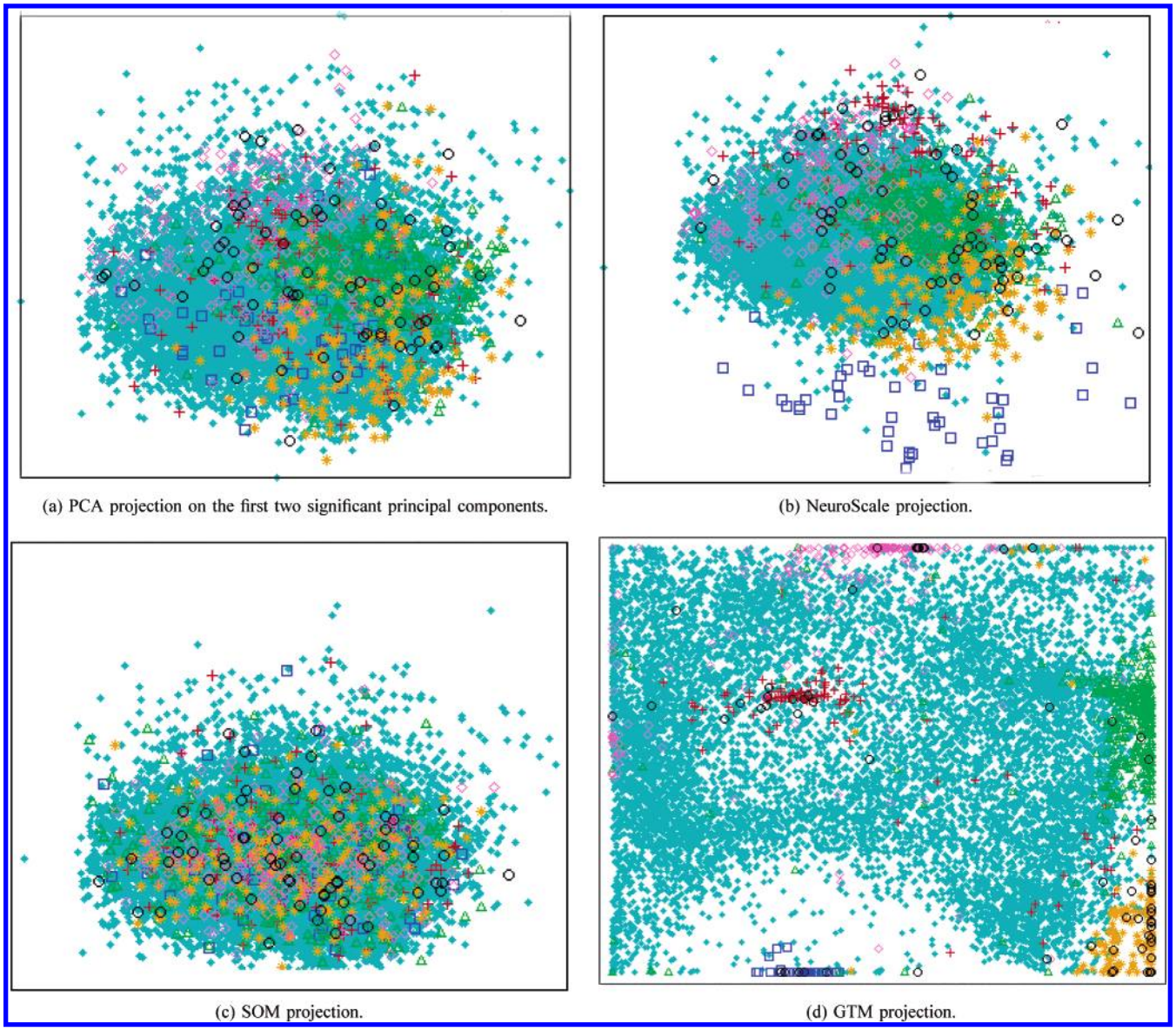

Figure 6. PCA, NeuroScale, SOM, and GTM projections. Refer to Table 1 for legend.

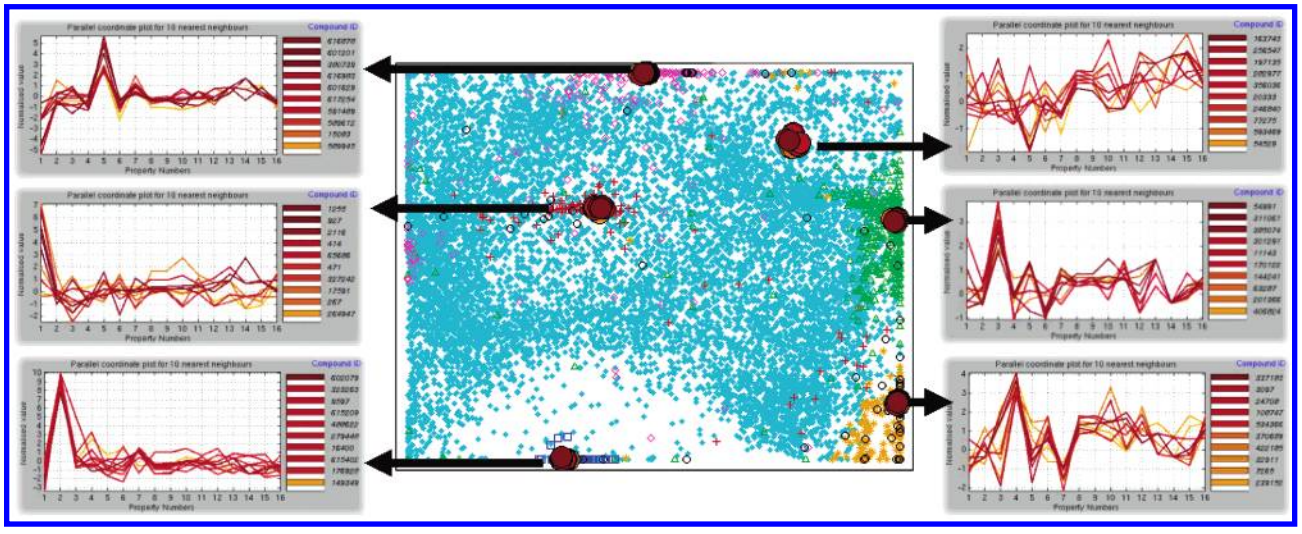

Figure 7. Local parallel coordinates demonstrating variations in the patterns in different regions of the GTM projection (plot 6d). Refer to Table 1 for legend.

where, because of the threshold, they are separated in a different bin.

Moreover, the corresponding magnification factor and directional curvature plots, for the GTM projection, presented

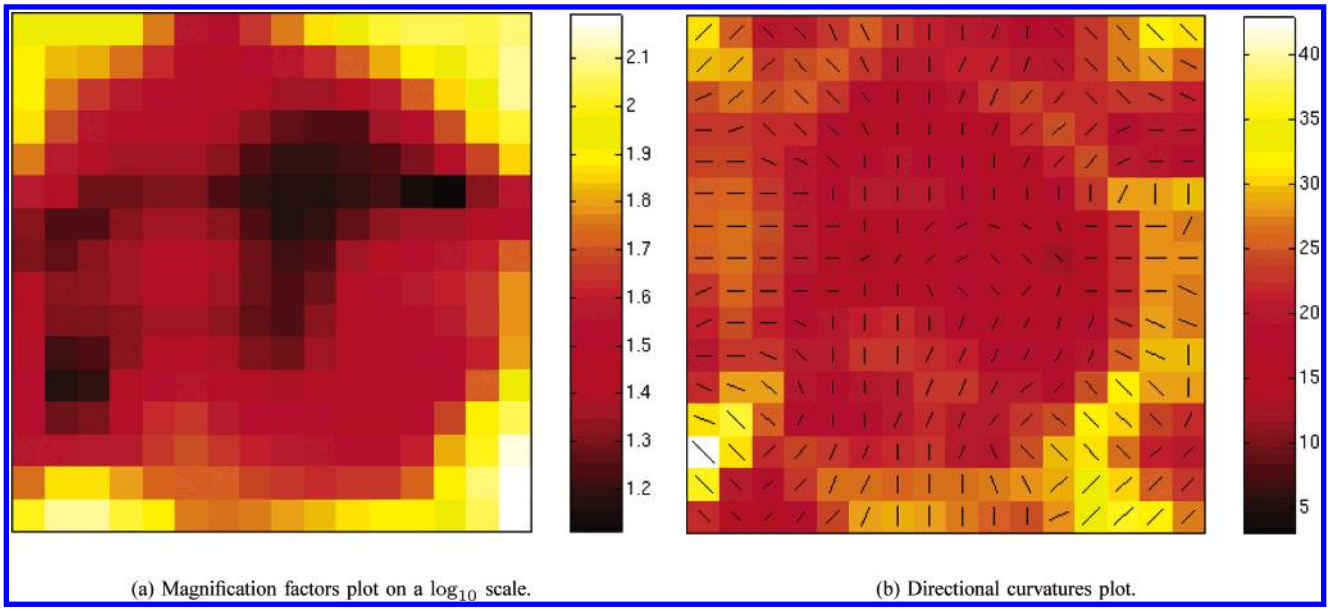

Figure 8. Magnification factor and directional curvature plots for the GTM projection (plot 6d).

in Figure 8, are useful to understand the structure of data in the data space. Using these plots, we can observe the stretching and folding of the projection manifold in the data space, respectively. The magnification factor plot is represented by color shading in the projection manifold. The lighter the color, the more the stretching in the projection manifold. The direction of folding in the projection manifold plot is presented using the direction line in the directional curvature plots. The length and the shade of the background color represent the amount of folding. The longer the line and the lighter the background color, the higher the folding (curvature).

Magnification factor and directional curvature plots are also useful for making decisions about the number and positions of the centers for GTM subplots during the training of an HGTM model. For example, the lighter bands at the bottom-right corner in the directional curvature plot (see Figure 8b) reveal a huge folding in the projection manifold to cover the data space. This helps us to understand that there could be a cluster there even though the data points are not marked (labeled) differently. Magnification factor and directional curvature plots are mainly used to understand projection space and data space in detail. But if the data are not colored (labeled; for example, if we do the analysis on a virtual compound library), magnification factor and directional curvature plots can be used to observe clusters in the projection and data space.

The HGTM visualization results are presented in Figure 9. Active compounds can be seen in different clusters in the root GTM. The deeper-level plots clearly separate interesting local regions. At each level, the magnification factor and directional curvature plots, presented in Figures 10 and 11, respectively, are used to make a decision about where to give the center of a subplot for the next level. Note that these results are only for 11 800 compounds (which is just about 1% of the total data set of 1 million compounds). The number of compounds one has to consider during the drug discovery process is enormous; in such situations, a well-trained HGTM model would be very useful to explore data at deeper levels.

Though visually we can easily observe the effectiveness of a GTM projection on this data set, it is useful to get an analytical measurement of the separation we obtained among different data classes in the projections. To obtain such a measurement of a projection result, first we fit a Gaussian mixture model (GMM)[21] on each class in the projection space and then we calculate the Kullback−Leibler (KL)[43] divergence among the fitted GMMs as below:

$$\mathrm{KL}(p_a,p_b) = \sum_x p_a(x) \log \frac{p_a(x)}{p_b(x)} \qquad (14)$$

where p_a and p_b are the GMMs for classes a and b, respectively.

The higher the value of the KL divergence, the greater the separation between the classes. Table 3 presents the sum of the entries in the KL divergence matrix for different visualization model outputs. GTM projection has a much higher KL divergence matrix sum, which is an evidence for better separation using GTM.

B. Application 2. It is also useful to visualize a data set with only the whole-molecular physicochemical properties as input to understand and prioritize the unscreened compounds for future HTS campaigns.

Figure 12 presents the PCA and GTM projections we obtained using the data set with only the 11 physicochemical properties described in section IV. Here, we can observe a soft grouping in the GTM projection (Figure 12b) for this data set compared to a blob we obtain in the PCA projection (Figure 12a). NeuroScale and SOM projection results are similar to those of PCA. The dominance of compounds active for different targets in different regions of the GTM projection (Figure 12b) could give us an opportunity to focus on the selected compounds. For example, the compounds active for the aminergic type 2 (marked as a green △) target are dominant in the top-right corner of the GTM projection (Figure 12b), so a careful study of compounds grouped in and near the top-right corner of the plot could be useful to prioritize compounds for the next HTS campaigns for the aminergic type 2 target.

Note that, in these plots, the biological activity data are merely used to label the data points on the plot for better

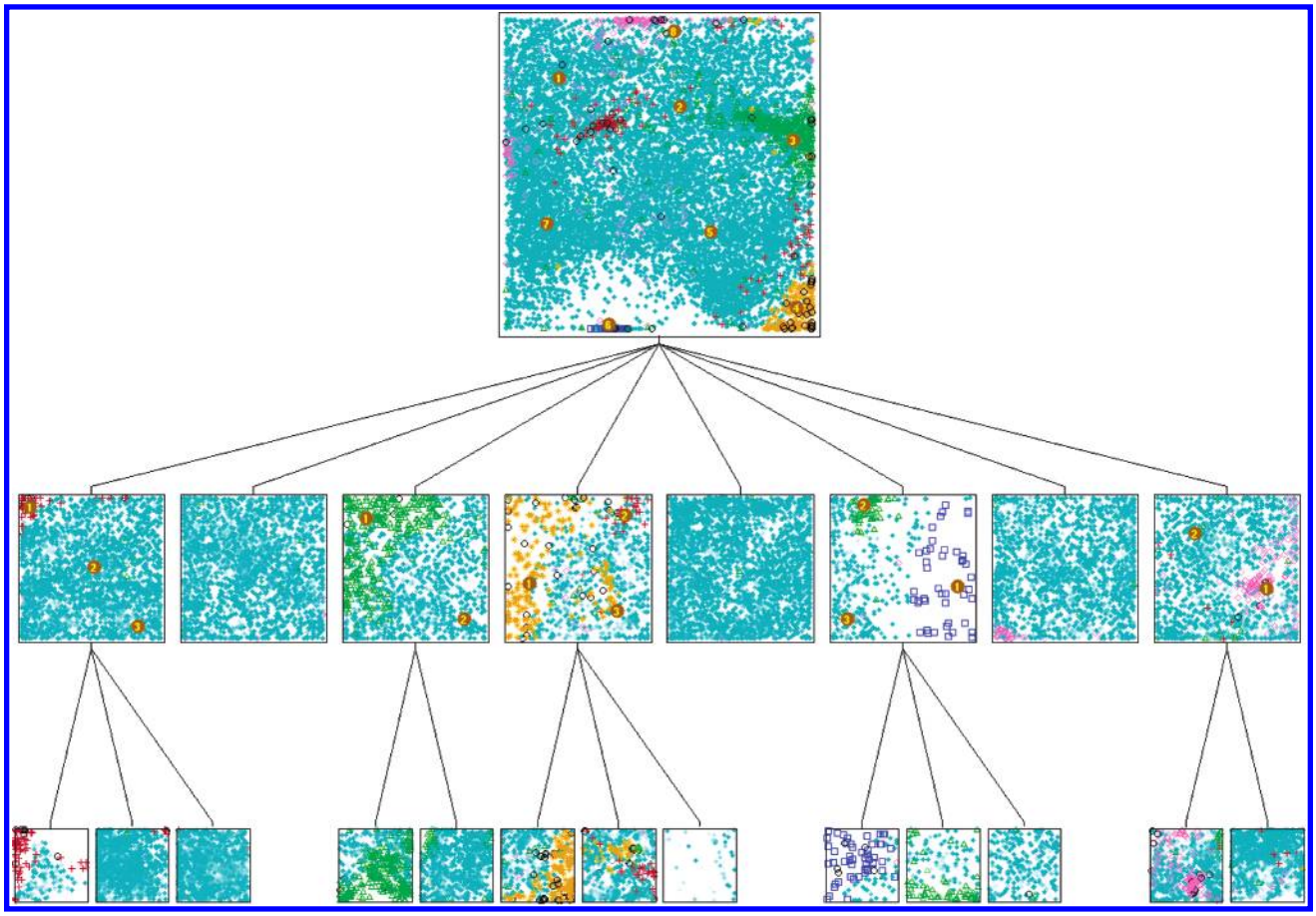

Figure 9. HGTM projection. Refer to Table 1 for legend.

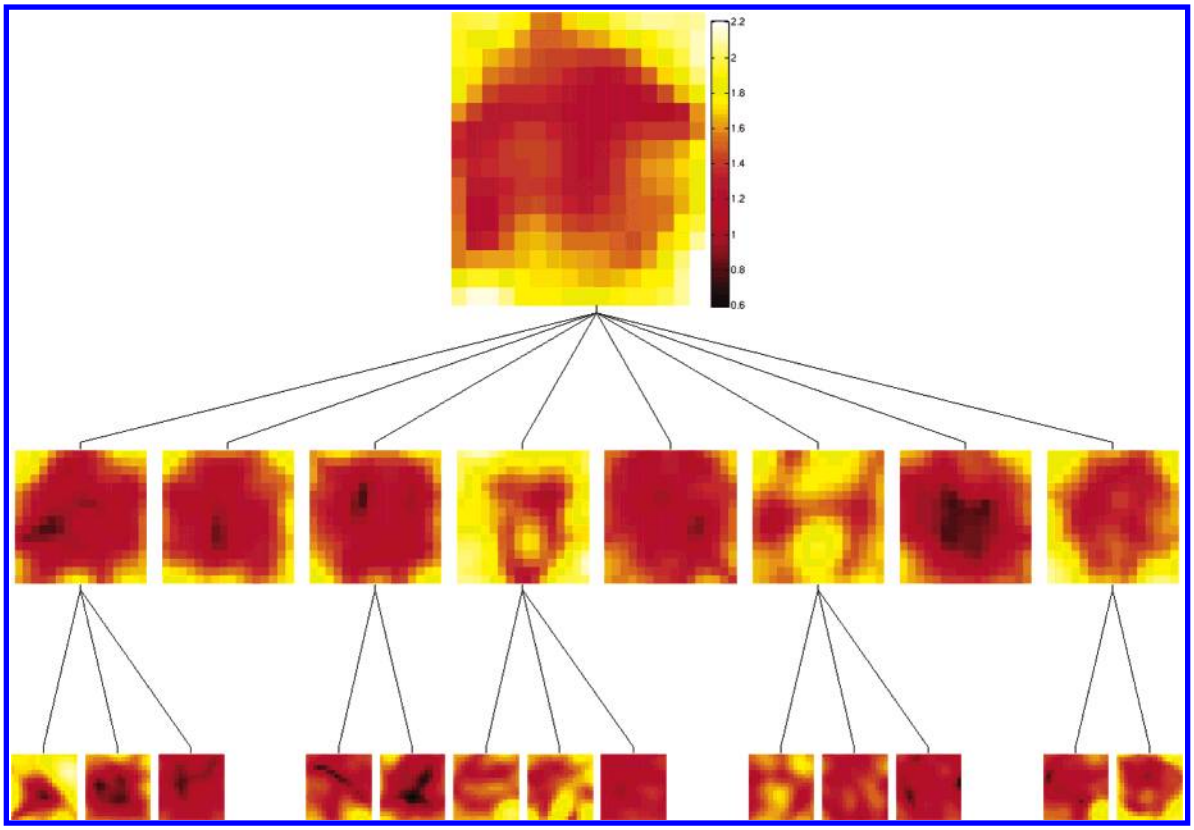

Figure 10. Magnification factor plot on a log scale for the HGTM projection (Figure 9).

presentation. Projections of only unscreened compounds from the virtual compound library may not have any labeling because if the compounds are from the virtual compound library, there may not be any biological activity data available

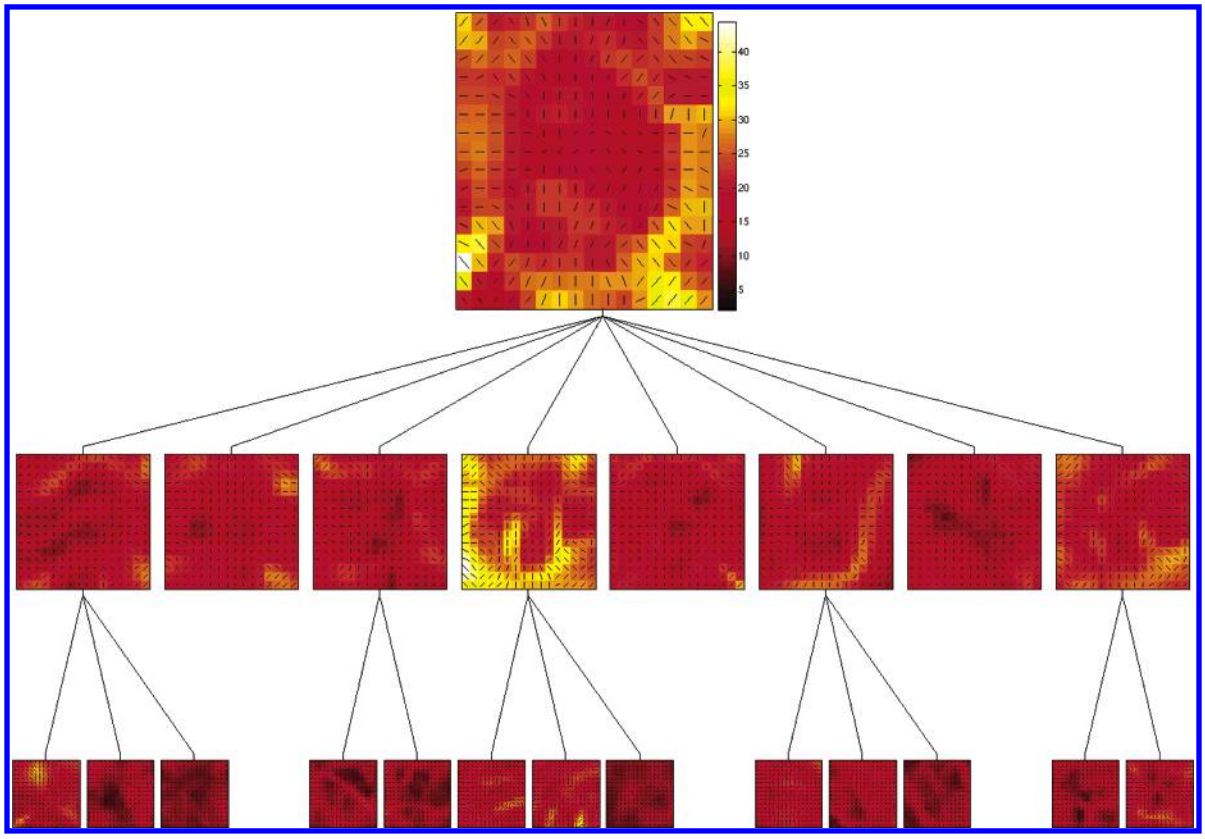

Figure 11. Directional curvature plot for the HGTM projection (Figure 9).

Table 3. KL Divergence Matrix Sum for Different Projection Models

visualization model	KL divergence matrix sum
PCA	33.63
NeuroScale	56.34
SOM	56.37
GTM	128.17

for them. Magnification factor and directional curvature plots for the GTM projection manifold, Figure 13, are useful in such a situation to understand the GTM projection manifold. It is also useful to add some already screened compounds in the data set so that we can easily locate regions of interest on the projection manifold by observing the positions of the screened active compounds on the projection manifold.

VI. DISCUSSION

The number of input dimensions, the structure of the data, and the quantity make it difficult to obtain a good projection using traditional visualization algorithms. An attempt to preserve relative similarities between the higher-dimensional input data and lower-dimensional projection space using PCA fails because of its linear nature. As we can see from the component matrix (Table 4), factor loadings of all biological activity variables are negligible for the first two significant PCs. The activity data is explained mostly by PCs 4 and 5. Figure 14 shows a projection using these two PCs. This projection provides a slightly better clustering of active compounds than the PCA projection obtained using the first two significant PCs (Figure 6a). The problem in doing so is that the physicochemical properties do not contribute greatly, so any analysis of inactive compounds adjacent to the actives is compromised.

Projection using Sammon's mapping (Neuroscale) gives us a blob because the outliers dominate the stress matrix used to fit the visualization model. The GTM first models the distribution of data in the data space, and then it gives us a uniformly distributed projection which provides a better clustering. A single 2D projection is not enough for a large data set because many points would be projected on top of others. Thus, the HGTM is useful to explore regions of a top-level GTM projection in detail.

One of our main aims in this research, to understand and explore biological activity data combined with other whole-molecular physicochemical properties (application 1, section V.A), was triggered from the need of the screening scientists at Pfizer to visually explore such a high-dimensional, large data set. The GTM and HGTM projections have proved very useful for the purpose and have outperformed the projections obtained from other visualization techniques. Even with only whole-molecular physicochemical properties as input during the model creation (application 2, section V.B), the GTM gave a relatively soft but useful grouping. As one can expect, the grouping in the projection obtained with only whole-molecular physicochemical properties as input is less clear than the grouping in the projection obtained with the biological activity data also being included in the input, because in the former case the model has less information to learn from. The performance could be improved by including more useful whole-molecular properties in the input data set (structure information, past HTS results on similar targets, etc.).

For GTM, the data distribution is not that important as (with enough kernels) it can model an arbitrary distribution.

Clearly, the leaf nodes of the HGTM hierarchy in Figure 9 represent individual groups of compounds. Using these

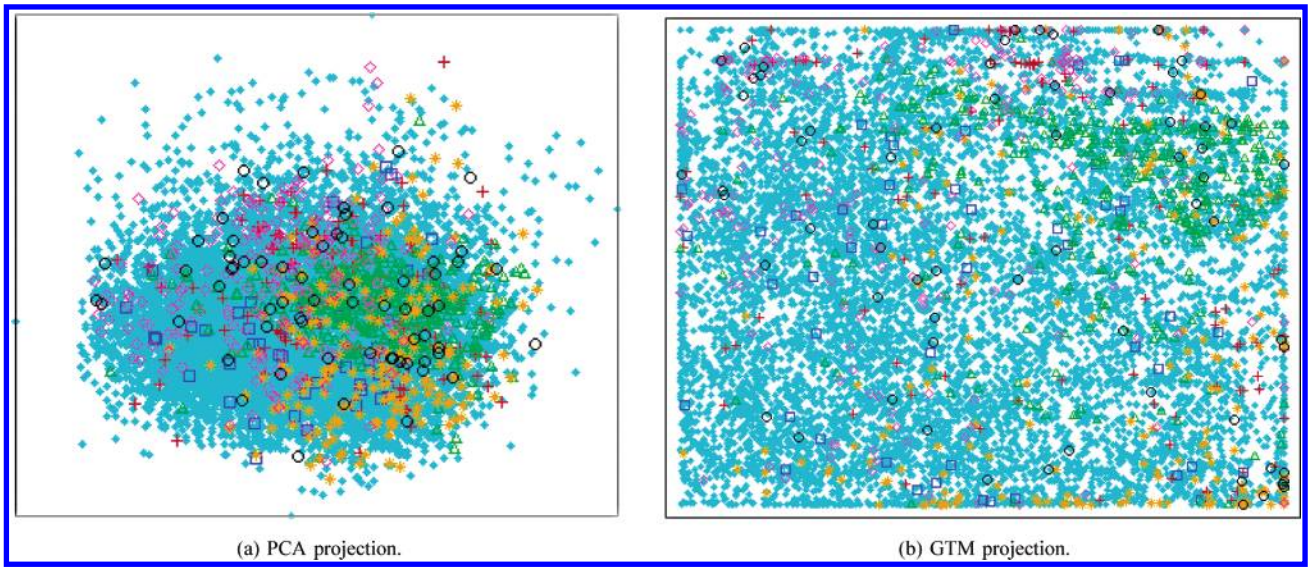

Figure 12. PCA and GTM projections using only the whole-molecular physicochemical properties. Refer to Table 1 for legend.

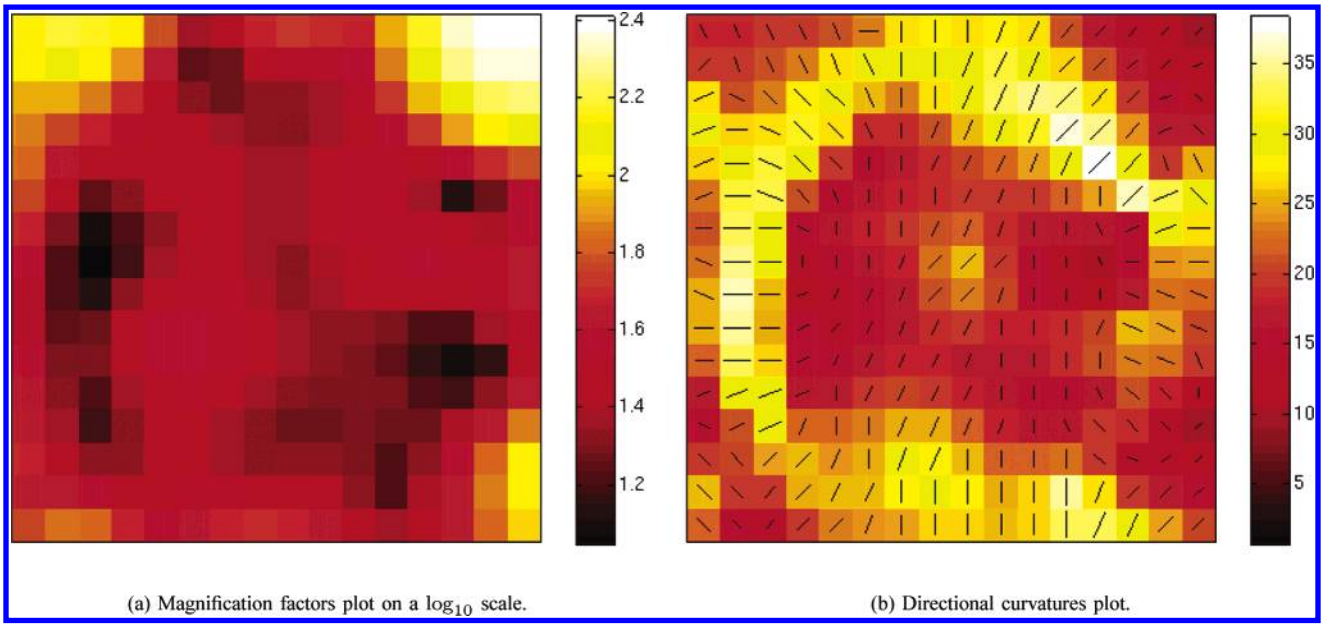

Figure 13. Magnification factor and directional curvature plots for the GTM projection (plot 12b).

groupings, it is possible to develop local classifiers to classify compounds for activity using physicochemical properties and past screening data. A single global classification/regression model can rarely capture the full behavioral variability of a huge multidimensional data set such as the one used here. Instead, local classification/regression (expert) models, each focused on a separate area of input space, often work better because the behaviors of different areas may vary. An effective grouping obtained using HGTM can be sensibly used to develop powerful guided local experts.[44]

VII. CONCLUSIONS

Deriving useful information from large-scale screening data is difficult because of the inherent noise and the sheer amount of data. Effective visualization is an important tool for chemists and biologists to understand large data sets. It enables them to link structural, physicochemical, and biological information with the biological activity of chemical compounds. Such understanding is proved to be useful to improve success rates for subsequent HTS campaigns.

Table 4. Component Matrix with Factor Loading Sorted in Ascending Order

	significant principal components (with eigenvalue > 1)				
	1	2	3	4	5
number of atoms	0.984				
number of bonds	0.974				
molecular weight	0.968				
number of hydrogens	0.757			0.327	
number of ring bonds	0.740	−0.310	0.310	−0.312	
molecular solubility	−0.726	0.478	0.301		
number of rotatable ring bonds	0.584	0.403	−0.268	0.411	−0.321
molecular polar surface area	0.393	0.781			
AlogP	0.511	−0.715			
number of hydrogen acceptors	0.517	0.684	0.268		
number of hydrogen donors		0.660	−0.395		
active for aminergic type 2			0.694		
active for aminergic type 1				0.561	
active for peptidergic type 2				0.392	0.389
active for kinase				−0.355	
active for peptidergic type 1					0.768

Traditional visualization techniques used in drug discovery, such as PCA, NeuroScale, and SOM, did not produce

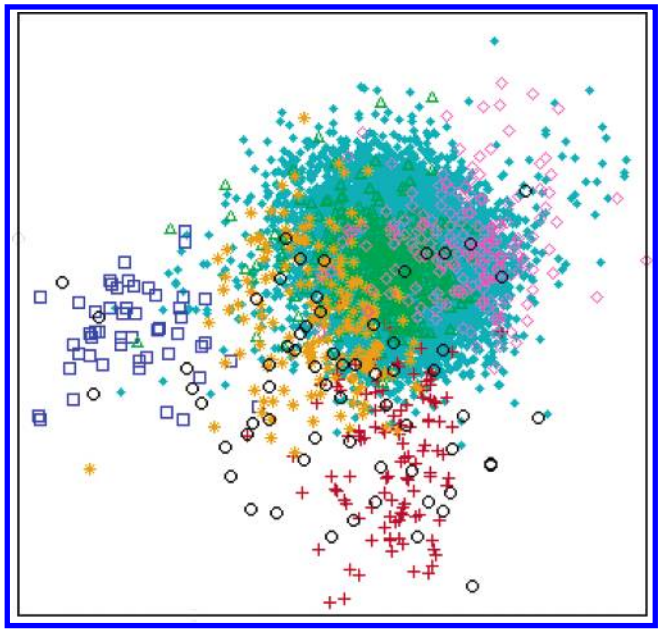

Figure 14. PCA projection on PCs 4 and 5. Refer to Table 1 for legend.

informative results for the data set used in this paper. These techniques proved to be ineffective and failed to generate additional knowledge, as we could not distinguish populations of molecules active for different biological targets.

For the data set analyzed here, GTM certainly gave much better results. The visualization results we obtained using GTM and the KL divergence matrix sum results presented in Table 3 clearly show the effectiveness of the clustering we obtain using GTM compared to that using PCA, NeuroScale, and SOM. The GTM and HGTM visualization results show a separation of active compounds for different screens. It not only enables us to characterize hit populations from different target classes (i.e., peptidergic GPCRs vs aminergic GPCRs vs kinases) but also enables us to understand areas of overlap. The observation that a few physicochemical parameters with a combination of screening results can be successfully used to characterize active from inactive compounds is in agreement with the recent work of Diller and Hobbs.[45] It is also in line with the observations at Pfizer that the properties (molecular weight, AlogP, polar surface area, etc.) of molecules active against different target classes show differences in the distribution and median of the property values.[46] Visualization allows the screening scientists to understand these relationships in relative to specific compounds.

One of the main advantages of GTM-based models is that it is possible to analytically calculate the magnification factors and the directional curvature of a GTM projection manifold. Magnification factor and directional curvature plots help to provide a better understanding of a projection manifold and, thus, the structure of the data in the data space.

An interactive software tool supporting these principled visualization algorithms and information visualization techniques has proved useful for a better understanding and exploration of the high-dimensional data sets. The integration of this tool with other industry standard software used in the drug discovery domain such as Pipeline Pilot and SpotFire gives the user more flexibility. Thus, the scientists can apply new algorithms such as GTM and HGTM while making best use of their existing software. Combining the effective use of all of this software and the use of principled visualization algorithms is likely to increase the odds in identifying active molecules and linking compound properties with biological activity.

ACKNOWLEDGMENT

D.M.M. is grateful to Pfizer Central Research and the Overseas Research Scholarship award for funding his Ph.D. The authors thank the anonymous reviewers for their comments, which have improved the manuscript.

REFERENCES AND NOTES

(1) Fox, S.; Farr-Jones, S.; Yund, M. A. High-throughput screening for drug discovery: Continually transitioning into new technologies. *J. Biomol. Screening* **1999**, *4*, 183−186.
(2) Handon, J. S. High-throughput screening − challenges for the future. *Drug Discovery World* **2002**, *Summer*, 47−50.
(3) Englebienne, P. High throughput screening: Will the past meet the future? *Front. Drug Des. Discovery* **2005**, *1*, 69−86.
(4) Bajorath, J. Rational drug discovery revisited: Interfacing experimental programs with bio- and chemo-informatics. *Drug Discovery Today* **2001**, *6*, 989−995.
(5) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.
(6) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.
(7) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(8) Bajorath, J. Integration of virtual screening and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(9) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(10) Oprea, T. I.; Zamora, I.; Ungell, A. L. Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.* **2002**, *4*, 258−266.
(11) Clark, E.; Pickett, S. D. Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49−58.
(12) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(13) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: Methods and models. *Drug Discovery Today* **2002**, *7*, S83−S88.
(14) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C. Selecting screening candidates for kinase and g protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256−1262.
(15) Schneider, G.; Nettekoven, M. Ligand-based combinatorial design of selective purinergic receptor (a2a) antagonists using self-organizing maps. *J. Comb. Chem.* **2003**, *5*, 233−237.
(16) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.
(17) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 647−679.
(18) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.
(19) Godden, J. W.; Xue, L.; Bajorath, J. Classification of biologically active compounds by median partitioning. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1263−1269.
(20) Bajorath, J. Virtual screening in drug discovery: Methods, expectations and reality. *Curr. Drug Discovery* **2002**, *March*, 24−28.
(21) Bishop, C. M. *Neural Networks for Pattern Recognition*, 1st ed.; Oxford University Press: New York, 1995.
(22) Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *C-18*, 401−409.
(23) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, 1995.
(24) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecule. *Curr. Med. Chem.* **2003**, *10*, 707−715.
(25) Balakin, K. V.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Ekins, S. Comprehensive computational assessment of ADME

properties using mapping techniques. *Curr. Drug Discovery Technol*. **2005**, *2* (2), 99−113.
(26) von Korff, M.; Steger, M. GPCR-tailored pharmacophore pattern recognition of small molecular ligands. *J. Chem. Inf. Comput. Sci*. **2004**, *44* (3), 1137−1147.
(27) Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G. Chemspace-shuttle: A tool for data mining in drug discovery by classification, projection, and 3d visualization. *QSAR Comb. Sci*. **2003**, *22* (5), 549−559.
(28) Lowe, D.; Tipping, M. E. Neuroscale: Novel topographic feature extraction with radial basis function networks. *Adv. Neural Inf. Proc. Syst.* **1997**, *9*, 543−549.
(29) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The generative topographic mapping. *Neural Comput.* **1998**, *10*, 215−234.
(30) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Magnification factors for the GTM algorithm. *Proceedings of the IEEE Fifth International Conference on Artificial Neural Networks*, July 7−9, 1997; pp 64−69.
(31) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Magnification factors for the som and gtm algorithms. *Workshop Proc. Self-Organizing Maps* **1997**, 333−338.
(32) Tiňo, P.; Nabney, I. T.; Sun, Y. Using directional curvatures to visualize folding patterns of the GTM projection manifolds. *Artificial Neural Networks − ICANN*; Dorffner, G., Bischof, H., Hornik, K., Eds.; Springer: Berlin, 2001; pp 421−428.
(33) Tiňo, P.; Nabney, I. T.; Sun, Y.; Williams, B. S. A principled approach to interactive hierarchical non-linear visualization of high-dimensional data. *Comput. Sci. Stat.* **2001**, *33*, 580−587.
(34) SciTegic. http://www.scitegic.com/ (accessed Oct 2005).
(35) Spotfire. http://www.spotfire.com/ (accessed Oct 2005).
(36) Tipping, M. E.; Lowe, D. Shadow targets: A novel algorithm for topographic projections by radial basis functions. *Neurocomputing* **1998**, *19*, 211−222.
(37) Tiňo, P.; Nabney, I. T. Constructing localized non-linear projection manifolds in a principled way: Hierarchical generative topographic mapping. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 639−656.
(38) Aurenhammer, F. Vornoi diagrams − survey of a fundamental geometric data structure. *ACM Comput. Surveys* **1991**, *3*, 345−405.
(39) Inselberg, A.; Dimsdale, B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. *Proc. IEEE Visualization* **1990**, 361−375.
(40) Levy, S. Interactive 3-d visualization of particle systems with partiview. *Proc. Int. Astronom. Union Symp.* **2001**, *208*, 85−91.
(41) MathWorks. http://www.mathworks.com/ (accessed Oct 2005).
(42) Nabney, I. T. *Netlab*: *Algorithms for Pattern Recognition*, 1st ed.; Springer: London, 2001.
(43) Cover, T.; Thomas, J. *Elements of Information Theory*, 1st ed.; Wiley: New York, 1991.
(44) Maniyar, D. M.; Nabney, I. T. Guiding local regression using visualisation. In *Deterministic and Statistical Methods in Machine Learning*, *LNAI*; Springer-Verlag: New York, 2005; pp 98−109.
(45) Diller, D. J.; Hobbs, D. W. Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem* **2004**, *47*, 6373−6383.
(46) Gribbon, P.; Sewing, A. High-throughput drug discovery: What can we expect from HTS? *Drug Discovery Today* **2005**, *10* (1), 17−22.

CI050471A