Journal of Molecular Graphics and Modelling 34 (2012) 108–117

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Topical perspectives

Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery

Michael Reutlinger, Gisbert Schneider*

Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 14 September 2011
Received in revised form
13 December 2011
Accepted 14 December 2011
Available online 2 January 2012

Keywords:
Chemical similarity
Drug design
Embedding
Molecular descriptor
Principal component analysis
Projection
Self-organizing map

ABSTRACT

Visualization of 'chemical space' and compound distributions has received much attraction by medicinal chemists as it may help to intuitively comprehend pharmaceutically relevant molecular features. It has been realized that for meaningful feature extraction from complex multivariate chemical data, such as compound libraries represented by many molecular descriptors, nonlinear projection techniques are required. Recent advances in machine-learning and artificial intelligence have resulted in a transfer of such methods to chemistry. We provide an overview of prominent visualization methods based on nonlinear dimensionality reduction, and highlight applications in drug discovery. Emphasis is on neural network techniques, kernel methods and stochastic embedding approaches, which have been successfully used for ligand-based virtual screening, SAR landscape analysis, combinatorial library design, and screening compound selection.

1. Introduction

The first modern atlas of the world, the "*Typvs Orbis Terrarvm*", was published in 1570 employing the Mercator projection of the globe (Fig. 1). There is no doubt that a two-dimensional (2D) map of the three-dimensional (3D) surface of the earth not only facilitated traveling from one place to the other, but more generally shaped our modern perception of the world. Similarly, it can be helpful to visualize chemical data in two dimensions, so that visual "navigation in chemical space" becomes possible. Visualization of compound distributions presents complex data in a simpler form [1,2]. By focusing on intrinsic dimensions of chemical data, relationships between compounds may be graphically displayed to inspire medicinal chemists and support hit finding and lead structure prioritization in drug discovery [3–5]. By "compound library" we here refer to a defined set of compounds, *e.g.* drug-like molecules or a combinatorial compound collection under investigation, rather than the whole universe of stable chemical structures. In this review, we present some of the essential mathematical concepts and motivate the use of nonlinear projection methods for vectorial numerical chemical data, which is obtained from representations of compounds by molecular descriptors like structural fingerprints, pharmacophoric features, or physicochemical properties. For an extensive overview of applications and practical examples of visualization in early drug discovery we refer to an excellent review article by Balakin and coworkers [4].

Often, the number of descriptors d used to encode molecular structure and properties exceeds the number of uncorrelated features by far, and dimensionality reduction and feature extraction methods are applied so that fewer "meaningful" descriptors or descriptor combinations are found. In other words, most multivariate compound data in $\mathbb{R}^d$ are not truly d-dimensional but form patterns on a lower-dimensional manifold [6]. In the context of this study, we refer to such a lower-dimensional molecular representation as a "projection" of data from a high-dimensional pattern space to a low-dimensional feature space $X \rightarrow X'$.

This concept of low-dimensional virtual screening and chemical data analysis is further motivated by several observations that can be made for high-dimensional chemical descriptor spaces [7]. With d approaching infinity, one encounters

Abbreviations: 2D, two-dimensional; 3D, three-dimensional; PCA, principle component analysis; PDB, Protein Data Bank; PPAR, peroxisome proliferator-activated receptor; QSAR, quantitative structure–activity relationship; SNE, stochastic neighbor embedding; SOM, self-organizing map; SPE, stochastic proximity embedding.

* Corresponding author at: Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, Wolfgang-Pauli-Str. 10, CH-8093 Zurich, Switzerland.
Tel.: +41 44 633 74 38/658 1616; fax: +41 44 633 13 79.
E-mail address: gisbert.schneider@pharma.ethz.ch (G. Schneider).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.12.006

Fig. 1. World Map "*Typvs Orbis Terrarvm*" (A. Ortelius, 1570). *Source*: The Library of Congress, Washington, DC, USA.

1. The *empty space phenomenon*: An exponential number of samples is needed to cover $\mathbb{R}^d$ in the sense that each dimension contains at least two compounds. In typical drug discovery scenarios, the chemical space spanned by a descriptor will be empty in terms of dataset coverage. For example, the maximum dimension that could be covered by a compound collection of one million compounds is as low as $\lfloor \log_2(10^6) \rfloor = 19$.
2. *Vanishing sphere volumes*: The volume of a d-dimensional Euclidean sphere with radius r becomes zero for $d \to \infty$. As a practical consequence for compound data from $\mathbb{R}^d$, there is a dimension d after which a sphere of radius r centered on compound x_i contains only x_i and no other sample. In other words, with increasing dimension of the molecular representation, the probability mass contained in a sphere with fixed radius around a compound decreases rapidly.
3. *Distance concentration*: Sample norms tend to concentrate and as a consequence, all distances are similar, samples lie on a hypersphere, and, each compound is nearest neighbor of all other compounds.

Consequences for virtual screening and the analysis of multivariate chemical data, including compound ranking and clustering, bear the danger of leading to erroneous results and consequently misinterpretation. We therefore motivate data visualization and dimensionality reduction methods as potentially very useful for hit finding and hit-to-lead optimization in early drug discovery. Specifically, we present and discuss principal component analysis (PCA), the encoder network, the self-organizing map (SOM), stochastic proximity embedding (SPE) and stochastic neighbor embedding (SNE) (Table 1).

2. Principal component analysis (PCA)

PCA is a linear dimension reduction method and belongs to the class of spectral dimension reduction methods. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of their variation (variance). This is achieved by transforming the data to a new set of uncorrelated variables, the *principal components* (PCs), which are ordered, such that the first few retain most of the variation present in all of the original variables [8]. PCs are linear combinations of the original descriptor axes and represent those directions in data space along which the scatter of the data is greatest. PCA has found widespread application in molecular modeling and drug design, and it is common practice to visualize compound distributions in graphical displays using the first two or three principal components [9].

PCA is performed by determining the eigenvectors and eigenvalues of the covariance matrix, or approximated values in case of large data matrices. The covariance of two random variables is their tendency to vary together. Suppose we have n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of a p-dimensional random variable (feature vector, molecular descriptor vector). The sample covariance matrix $\mathbf{S}$ is given by Eq. (1), with $\bar{\mathbf{x}}$ being the sample mean and the superscript T denotes the matrix transpose.

$$\mathbf{S} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^{\mathrm{T}} \tag{1}$$

In the case of centered data with $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ the covariance matrix is rewritten (Eq. (2)):

$$\tilde{\mathbf{S}} = \frac{1}{n-1} \sum_{k=1}^{n} \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^{\mathrm{T}} \tag{2}$$

By solving the covariance matrix $\tilde{\mathbf{S}}$ for eigenvectors $\mathbf{a} \in \mathbb{R}^p$ and eigenvalues $\lambda \in \mathbb{R}_{\geq 0}$, subject to the constraint that $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$, we can find the projection directions (Eq. (3)).

$$\tilde{\mathbf{S}}\mathbf{a} = \lambda \mathbf{a} \tag{3}$$

Table 1
Comparison of selected projection methods.

	PCA	Encoder network	SOM	SPE	SNE
Principle of the projection	Linear	Nonlinear	Nonlinear	Nonlinear	Nonlinear
Preserved property	Variance	Identity	Neighborhood	Distance	Neighborhood probability
Bijective projection	Yes	Limited, potentially infinite number of solutions	Limited to neuron vectors	No	No
Out-of-sample projection of new data	Yes	Yes	Retraining required	Retraining required	Retraining required
Computational demand	Low	High	Medium	Low	High
Software availability	High	Low	High	Low	Low

Data points $\tilde{\mathbf{x}}_i$ are transformed into the new PC coordinate system by orthogonal projection of the data points on each of the eigenvectors (Eq. (4)).

$$\tilde{\mathbf{z}}_i = \tilde{\mathbf{x}}_i \mathbf{A} \tag{4}$$

where $\mathbf{A}$ is the orthogonal $(p \times p)$ matrix with the eigenvectors as columns. Eigenvectors are sorted by decreasing eigenvalues, which can be interpreted as their "significance". To obtain a lower dimensional projection of the original data eigenvectors with small eigenvalues are omitted from $\mathbf{A}$.

Ten years ago, Oprea and Gottfries presented the ChemGPS (chemical global positioning system) in combination with PCA for the linear projection of chemical data [10]. Its main feature is a set of "satellite" compounds that are placed outside druglike space and thereby define outer data borders, and consequently, the applicability domain of the projection. The original application employed a total of 423 satellites and representative drugs ("core structures"). This concept of defining the borders of a chemical space by extreme-valued compounds can help identify projection artifacts and prevent unjustified conclusions from inspection and interpretation of chemical space maps and is not limited to PCA. In fact, it is recommended to use a basis set of reference cores and satellites for any projection of compound distributions. With the advent of large open access repositories and searchable databases of bioactive compounds – *e.g.* ChEMBL [11], PubChem [12], ChemBank [13], ChEBI [14], ChemDB [15] – the known bioactive chemical space is continuously extended and refined [16]. This huge body of chemical structures and literature data will help in defining appropriate boundaries of druglike chemical space [17].

Despite its appeal there are certain limitations of PCA that motivate nonlinear projection techniques to be used complementarily, *e.g.* the requirement for normal-distributed data, susceptibility to outliers, and issues with data manifolds and large data sets, to just name some prominent examples. In drug discovery one is mainly interested in the structure of local neighborhoods of known bioactive compounds or reference molecules [18], and the *Chemical Similarity Principle* [19] is grounded on the neighborhood concept [20,21]. PCA *per se* does not preserve local structure of the input data in the projection. In contrast, nonlinear embedding techniques do not assume global linearity but make a weaker local linearity assumption. In high-dimensional input space the Euclidian (L_2 norm) distance is assumed to be a good measure of geodesic distance (*vide infra*) only for nearby points, which is also observed for chemical data suffering from the "curse of dimensionality" [22].

We often face nonlinearity in structure–activity relationship (SAR) modeling, which manifest as perceived "activity cliffs", that is, when structurally similar (nearby) compounds exhibit significantly different pharmacological or other measured effects. Seemingly, the *Chemical Similarity Principle* does not hold in these regions of chemical space. This assumption may not be true, as the measured effect of a considered small change of chemical structure, *e.g.* an exchange of methyl by ethyl, can be dramatic if an essential, function-determining molecular feature (pharmacophore) is destroyed [23,24]. In order to account for nonlinearity, specifically its relation to some observable function or property, one can either conceive appropriate, context-dependent molecular descriptors that restore global linearity, or apply nonlinear, local neighborhood preserving embedding methods that are able to capture manifolds in high-dimensional chemical data [25].

Numerous nonlinear projection methods – expressly manifold learning techniques such as Local Linear Embedding (LLE) [26], the IsoMap [27] approach and its derivatives Laplacian [28] and Kernel IsoMap [29] – have found widespread application in natural sciences, in particular bioinformatics [30–33], but mainly outside of chemistry. Some of these methods may be considered as specific instances of Kernel PCA [34,35], which employs the "kernel trick" to perform conventional linear PCA not in the original input space X (*i.e.* the original molecular descriptors), but in a virtual, very high-dimensional Hilbert space V, so that nonlinear relationships in X will gain linear meaning in V. In other words, the kernel trick virtually increases the dimensionality of the input data so that, in this higher dimension space, they become linearly related. The particular appeal of kernel methods is that the Hilbert space is never explicitly generated by transforming the original data into that space, but implicitly computed using a kernel function Φ. The Support Vector Machine (SVM) represents a prominent machine-learning concept using the kernel trick. SVMs are most successfully employed for SAR modeling and classification of chemical data [36]. More recently, kernel-based Gaussian process modeling has been introduced to chemistry and drug discovery [37,38]. While kernel methods provide an elegant approach for nonlinear SAR modeling and classification, they do not explicitly provide a means for data visualization and interpretation of complex nonlinear models. With few exceptions of chemical feature extraction, visualization and interpretation published [39–41], this might be a reason why kernel techniques have not found excessive appreciation in medicinal chemistry yet.

3. Encoder network

Special types of feed-forward artificial networks were among the first nonlinear methods employed for dimensionality reduction in chemistry [42–44]. One such system of particular interest is the symmetric encoder network (Fig. 2) [45,46]. Here, the idea is to simultaneously compute a nonlinear forward- and a back-projection of chemical data. Within the applicability domain and under certain conditions [47], this concept would allow one to navigate in the low-dimensional projection, and for each position of this map the coordinates (substructures, properties, or pharmacophoric features – depending on the molecular descriptor used) of compound in the original space are provided, thereby solving the "inverse QSAR problem" [48,49]. Despite its appeal only few applications of the encoder network approach have been published. One reason for its limited use might be the small number of ready-to-use software tools implementing such a system (one such free tool is the software *ChemSpaceShuttle* [50]). One also needs to take into account long training times and the requirement for an optimization of network architecture, specifically the number of hidden neurons.

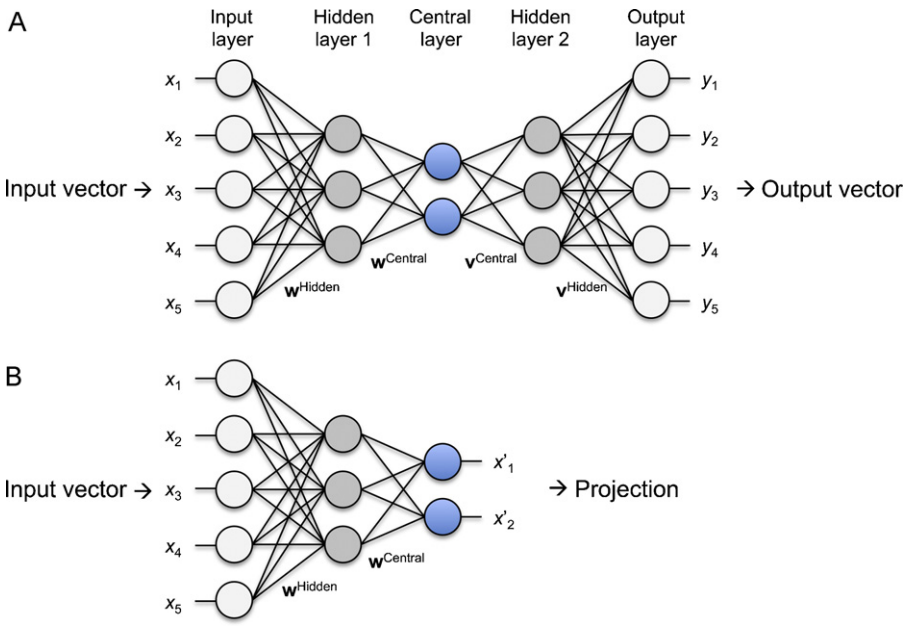

Fig. 2. Topology of an encoder network for projection of multivariate data. Artificial neurons are drawn as circles, connection weights as lines between neurons in the different network layers. The symmetric network is trained so that any output vector **y** ideally is identical to its corresponding input vector **x** (A). After successful network training the central layer neurons (blue circles) compute the desired projection of **x** (B). The network structure shown accepts a five-dimensional input vector and computes a two-dimensional projection **x**′. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

An encoder network must be trained in a supervised fashion, *i.e.* a forward (encoder) and a backward (decoder) function must be found, by using sets of reference compounds represented by a vectorial molecular descriptor **x**. Having defined the number of hidden neurons and the desired projection (number of central layer neurons), the weights of the encoder–decoder function defined by the network's architecture ($\mathbf{w}^{\text{Hidden}}$, $\mathbf{w}^{\text{Central}}$, $\mathbf{v}^{\text{Hidden}}$, $\mathbf{v}^{\text{Central}}$) are optimized so that for every input vector **x**, an identical output vector **y** is computed (Fig. 2A).

Typically, variations of the delta rule in combination with gradient-based or stochastic optimizers are employed for minimizing $\mathbf{x} - \mathbf{y}$. After successful training, the network "compresses" the input data when smaller numbers of neurons are used in the central layer than in the input layer, and can be used for generating projections of **x** as outputs of the central layer neurons **x**′ (Fig. 2B). Eq. (5) gives the simplified network function computing the actual projection, where ϑ and $\boldsymbol{\theta}$ are hidden and central neuron bias values.

$$x'_i = f(\mathbf{x}) = \sum_{j}^{\text{Hidden}} \left(w_j^{\text{Central}} \left(\sum_{k}^{\text{Input}} w_k^{\text{Hidden}} x_k \right) + \vartheta_k \right) + \theta_i \qquad (5)$$

It must be kept in mind that there is an infinite number of input vectors that are projected to the same point in **x**′, and therefore, particular care must be taken to clearly define the applicability domain of encoder network-based QSAR models [51,52]. Only recently, this topic has been re-visited and appropriate techniques for applicability domain estimation for various machine learning models have been proposed [53–55].

4. Self-organizing map (SOM)

The concept of self-organizing mapping of high-dimensional input was conceived by Kohonen in the early 1980s [56], and introduced to chemistry and drug design by Gasteiger and co-workers in the 1990s. The SOM (or "Kohonen net") has been extensively applied in drug discovery ever since [57,58]. The SOM architecture consists of a regular array of so-called "neurons", which essentially are vectors that are arranged in a topological structure (typically a 2D array) and have the same dimension as the input data. During the SOM training process – an optimization procedure following the principles of unsupervised, associative Hebbian learning [59] – the original high-dimensional space is tessellated, resulting in as many data clusters as there are neurons in the SOM. The neurons represent centroids of each cluster (Voronoi field). Data points within a cluster are more similar to "their" neuron than to any other neuron of the SOM. In this regard, SOM training may be considered a variant of *k*-means clustering, similar to vector quantization [60,61]. The resulting prototype vectors capture features in the input space that are unique for each data cluster. Molecular feature analysis can be done, *e.g.*, by comparing adjacent neuron vectors. In analogy to the Mercator projection shown in Fig. 1, Fig. 3A presents a SOM projection from 3D coordinates of points on the earth's surface to a 2D map. The SOM grid contains 2400 neurons arranged as a toroidal 60×40 grid. Some major city locations are highlighted as reference points. Note that although the overall distribution and shapes of continents and oceans is not metric, local neighborhoods are preserved, *e.g.* London and Zurich are close to each other on the 2D map.

Kohonen's algorithm represents an efficient way for mapping vectors that are close to each other in input space onto contiguous locations in the output space. Preservation of *local* neighborhood is achieved by introducing a topology to the layout of the SOM neurons. The simplest topology is a chain of neurons, followed by a 2D grid. Molecules that are located in adjacent clusters on the neuron grid are also close to each other in the original high-dimensional space. Topological mapping can be achieved by two simple rules that guide the training process:

1. For input vector **x** locate the best-matching neuron ("winner" neuron, $\mathbf{w}^*$).
2. Move this neuron and its topological neighbors toward **x**.

For the first rule vector distances between **x** and the SOM neurons **w** have to be computed (Eq. (6)). The number of comparisons

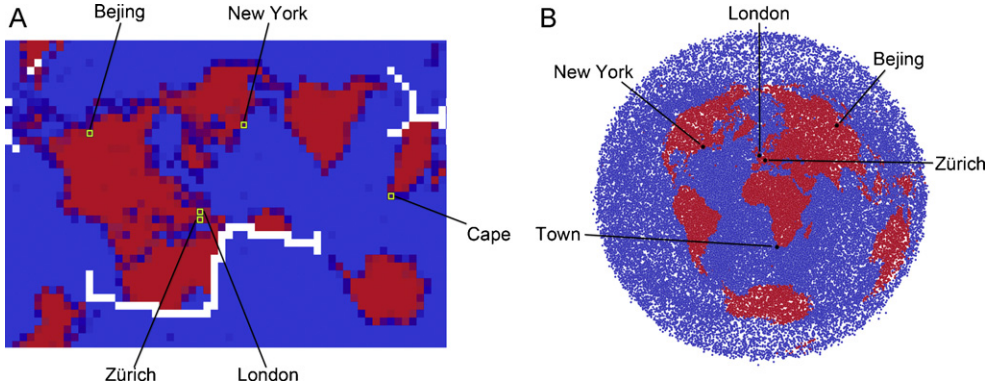

Fig. 3. Projections of the land-water distribution on the surface of the earth (*red*: land, *blue*: water). The left panel (A) presents a 2D SOM (60 × 40 neurons) projection of these data. In (B) ISPE ($r_c = 0.2$) was used to generate a 2D map. White color in (A) indicates "empty space", *i.e.* neurons without data points assigned. Earth data were obtained from the NASA NEO data set "Blue Marble: Next Generation (Terra/MODIS)". The 2D latitude/longitude data, as defined by the position in the image, were transformed to Cartesian coordinates. Random sampling from the sphere surface was applied to select subsets of 15,007 (SOM) and 65,167 points (ISPE), equally distributed over the sphere. The binary land/water descriptor was calculated from the color information in HSV color space.

needed depends linearly on the size of the self-organizing system *S*, which can be expressed by its number of neurons.

$$||\mathbf{x} - \mathbf{w}_i|| \rightarrow \min. (\forall i \in S) \quad (6)$$

The second rule requires an updating procedure to adapt the vector elements of the winner neuron $\mathbf{w}^*$ and its topological neighbors (Eq. (7)), where ε is a learning rate depending on both the topological distance between $\mathbf{w}^*$ and neuron $\mathbf{w}_i$, and on the training time passed. A toroidal neuron topology can be used to avoid some boundary problems inherent to a planar topology. For a full description of the SOM algorithm see the literature [62].

$$\mathbf{w}_i = \mathbf{w}_i + \varepsilon||\mathbf{x} - \mathbf{w}_i|| \quad (7)$$

Fig. 4 presents an application of SOM-based virtual screening for new kinase inhibitors [63]. The idea in this study was to map and cluster known drugs and lead compounds and a virtual combinatorial library on a 2D SOM. All compounds were represented by 150-dimensional CATS descriptors, a topological pharmacophore feature representation [64,65]. The SOM was then colored according to the prevalence of combinatorial compounds (Fig. 4A) and known kinase inhibitors (COBRA data collection, Fig. 4B). The neuron containing the most combinatorial compounds coincides with a kinase "activity island" [66]. This particular cluster holds many of the reference inhibitors (seven-fold overrepresentation of kinase inhibitors compare to the background distribution), and it was therefore reasonable to assume similar targets for the combinatorial compounds. One candidate (compound **1**) from the virtual library was actually synthesized and successfully tested in a CDK2 inhibition assay.

The SOM virtual screening approach presented in Fig. 4 belongs to the class of ligand-based similarity searching methods [67,68]. In contrast to using reference compounds as queries and ranking the combinatorial screening compounds by some pharmacophore similarity index, the SOM offers the potential advantage of performing similarity searching using a "common pharmacophore" model (*i.e.* the neuron vector) as query. This avoids the necessity for comparing and merging ranked lists of candidate compounds [69]. Despite its appeal, the SOM approach used in this study has several disadvantages compared to other ligand-based virtual screening techniques. A major limitation of the original SOM algorithm is that the dimension of the output space and the number of neurons must be predefined prior to SOM training [70]. A disadvantage of SOMs can be the comparatively long training time needed, especially if large data sets (*e.g.* HTS screening data, combinatorial compound libraries) are used. Different training runs bear the additional danger of delivering slightly different results due to the stochastic nature of SOM optimization. Several variations and extensions of Kohonen's original SOM algorithm have been published and applied to drug discovery [71]. Such developments include self-organizing networks with an adapting grid size [72], cascaded SOMs [73], and hybrid neural networks [74,75]. These systems might provide alternative approaches to virtual compound screening, although their practical usefulness and applicability to hit and lead finding still needs to be rigorously assessed.

5. Stochastic proximity embedding (SPE)

SPE is a self-organizing dimensionality reduction algorithm that aims at preserving the pairwise proximities in the lower dimensional embedding. It was introduced to drug discovery by Agrafiotis and coworkers in 2002 [76]. In its first version, SPE was used to find a stochastic approximation of multidimensional scaling that preserved the metric structure. The mapping procedure uses a pairwise refinement strategy that does not require the complete distance or proximity matrix and scales linear with the size of the data set. This allows dimension reduction even for large data sets.

If the data forms a lower dimensional nonlinear manifold, conventional similarity measures, such as the Euclidean distance, tend to underestimate the proximity of points and lead to erroneous embedding. To properly reconstruct the manifold, the geodesic distance, the proximity of two points measured on the manifold itself, needs to be preserved. Algorithms like IsoMap [26] or LLE [27] estimate the geodesic distance from the local neighborhood. Agrafiotis observed that the geodesic distance is always greater or equal to the input proximity. If two points are close, the input proximity provides a good approximation to their geodesic distance; when they are further away, the input proximity provides a lower bound [77]. Isometric SPE (ISPE) circumvents the calculation of estimated geodesic distances by incorporating this observation [6]. It forces embedding distances of nearby points to match their input proximities, while points whose input proximities are larger than a defined threshold are forced to stay further apart. ISPE preserves the distances of the local neighborhood and views the distances between remote points as lower bounds of their true geodesic distances and uses them to impose the global structure (Fig. 3B). A similar approach has been applied to SOM training, where neurons adjacent to the winner neuron $\mathbf{w}^*$ are attracted to the data point presented, while neurons outside a defined topological neighborhood on the SOM grid are pushed away. Such a procedure can also be used to enhance contrast on SOMs [78].

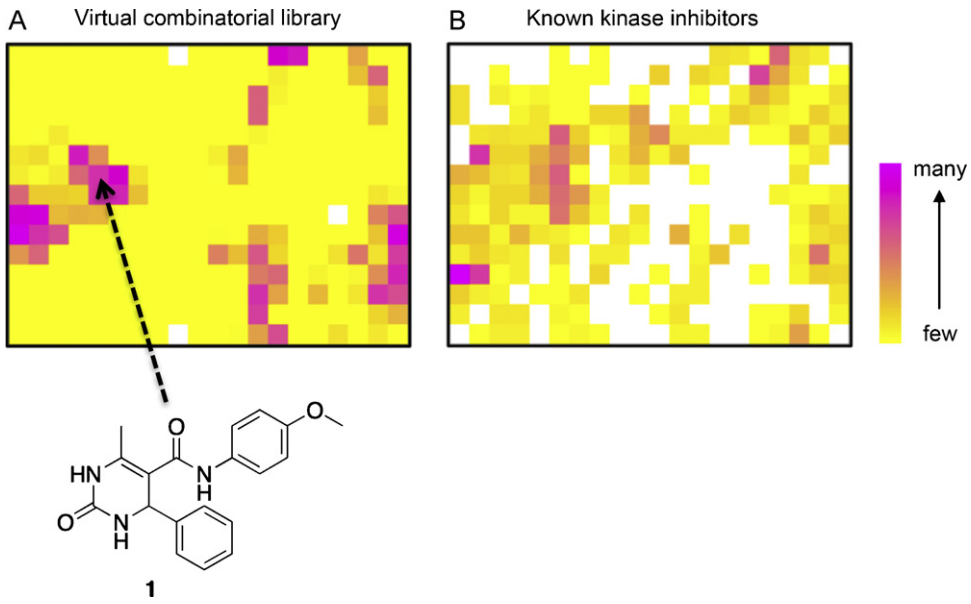

Fig. 4. 2D SOM projection of a virtual combinatorial library containing Biginelli-type dihydropyrimidones (A) and drug-like bioactive compounds with kinase inhibitors highlighted (B). The SOM grid contains (15 × 20) neurons. By SOM analysis, compound **1** was identified as an inhibitor of cyclin-dependent kinase 2 (CDK2).

The stress function S minimized stochastically by ISPE is given by Eq. (8).

$$S = \frac{\sum_{i<j}(f(d_{ij}, r_{ij}))/r_{ij}}{\sum_{i<j} r_{ij}} \rightarrow \text{min.}, \tag{8}$$

where r_{ij} is the input proximity between the ith and jth point, d_{ij} is their Euclidean distance in the low-dimensional embedding space and $f(d_{ij}, r_{ij})$ is the pairwise stress function defined as $f(d_{ij}, r_{ij}) = (d_{ij} - r_{ij})^2$ if $r_{ij} \leq r_c$ or $d_{ij} < r_{ij}$ and $f(d_{ij}, r_{ij}) = 0$ if $r_{ij} > r_c$ and $d_{ij} \geq r_{ij}$. r_c is the neighborhood radius.

ISPE minimizes the stress function with a stochastic steepest descent approach. It iteratively refines a starting configuration of the data points by repeatedly selecting two points at random and adjusting their coordinates so that their embedding distance d_{ij} matches more closely their input proximity r_{ij}. The correction is proportional to the disparity $\lambda|r_{ij} - d_{ij}|/d_{ij}$, where λ is a learning rate. To avoid oscillation the learning rate is decreased during the course of refinement. If the points are not neighbors, if $r_{ij} > r_c$ and $d_{ij} > r_{ij}$, their coordinates remain unmodified.

The result of ISPE strongly depends on the choice of the neighborhood radius for learning the embedded manifold. If r_c is too large, shortcuts to other branches of the manifold are possible, whereas if it is too small it may lead to fragmented clusters (Fig. 5).

As an illustrative example, Fig. 6 presents the application of PCA and ISPE to producing a 2D chemical structure depiction from 3D molecular atom coordinates. The first compound is PPAR-gamma agonist pioglitazone, for which a receptor-bound conformation served as "high-dimensional" input. Apparently, both PCA and ISPE are able to produce a 2D visualization lacking great distortion. This can be explained by the extended shape of the pioglitzone conformer, which is properly projected on the first two PCs. The second example provides 2D projections computed for epothilone D bound to cytochrome P450epoK. Here, ISPE generates a more appealing, less compact 2D mapping than PCA, probably due to the greater degree of 3D conformational folding of the reference structure.

6. Stochastic neighbor embedding (SNE)

In contrast to SPE, SNE does not try to preserve pairwise *distances* but instead the *probabilities* of points being neighbors [79]. The pairwise distances in the input and output space are used to calculate the probability distributions that point i is a neighbor of point j. The aim of the embedding is to approximate the neighbor probability distribution as close as possible in the low-dimensional embedding.

The probability of point i being neighbor to point j in the input space is defined as (Eq. (9)):

$$p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \tag{9}$$

The proximities d_{ij}^2 may be given as a proximity matrix or calculated using the scaled squared Euclidean distance between the high-dimensional input data points $\mathbf{x}_i$ and $\mathbf{x}_j$ (Eq. (10)).

$$d_{ij}^2 = \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2} \tag{10}$$

The scaling factor σ_i, the variance of the Gaussian at point $\mathbf{x}_i$, is either set manually or by fixing the entropy of the distribution. Setting the entropy to $\log k$ sets the "effective number of local neighbors" to k [79].

The induced probability q_{ij} that point i picks point j as its neighbor in the embedding space, with $\mathbf{y}_i$ being the low-dimensional images, is defined as (Eq. (11)):

$$q_{ij} = \frac{\exp(-||\mathbf{y}_i - \mathbf{y}_j||^2)}{\sum_{k \neq i} \exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)}. \tag{11}$$

The aim is to match the probabilities as close as possible, as measured by the sum of Kullback–Leibler (KL) divergences between the original and induced distributions over neighbors for each object (Eq. (12)).

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i||Q_i) \tag{12}$$

Hinton et al. already showed that the probabilistic framework can easily be extended to allow multiple low-dimensional images for each high-dimensional object through a mixture of Gaussians [79]. In the past few years several extensions to the classical SNE algorithm have been published. T-distributed stochastic neighbor embedding (t-SNE) [80] uses a symmetric cost function, which is

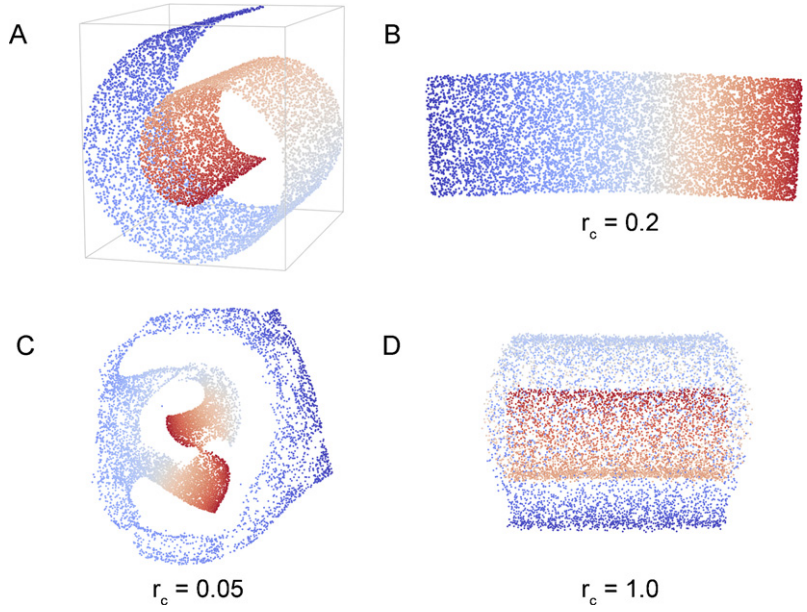

Fig. 5. ISPE projections of the 3D "Swiss roll" manifold (A) to two dimensions (B–D) using different cut-off distances r_c. The data points for the Swiss roll were obtained by generating coordinate triplets $\{x = i\cos(i), y = i\sin(i), j\}$, where i and j are random numbers from the intervals [5, 15] and [0, 30], respectively. The color corresponds to the angle i. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

easier to optimize, and a Student-t distribution instead of the Gaussian to model similarity in low-dimensional space. It improves visualization because natural clusters tend to be more separated in the low-dimensional embedding, thus it simplifies the visual perception of clusters. The application was extended to visualize data together with class labels [81], incorporate multiple similarity matrices [82], or multiple views of the input data [83]. SNE has also been analyzed within the framework of information retrieval and a variant has been introduced, which optimizes the retrieval quality, quantified by precision and recall [84]. A limitation of SNE is its computational demand for projecting large datasets due to the calculation of the complete pairwise probability matrix and steepest-decent optimization. To some extent this has been addressed in recent publications using trust-regions to speed-up convergence [85], or landmark sampling to reduce memory consumption [79].

A typical application of SNE is the visualization of the distribution of combinatorial compound libraries in some high-dimensional pattern space spanned by pharmacophore descriptors (Fig. 7). As an example, we encoded a library of 15,840 three-component Ugi reaction products by the topological CATS descriptor. Fig. 7A presents a 2D projection of these 250-dimensional data that was obtained by ISPE. Color corresponds to measure inhibitory activity of the compound against tryptase. An "activity island" containing many inhibitors is highlighted in red color (low IC_{50} values). Visualization compound distributions can

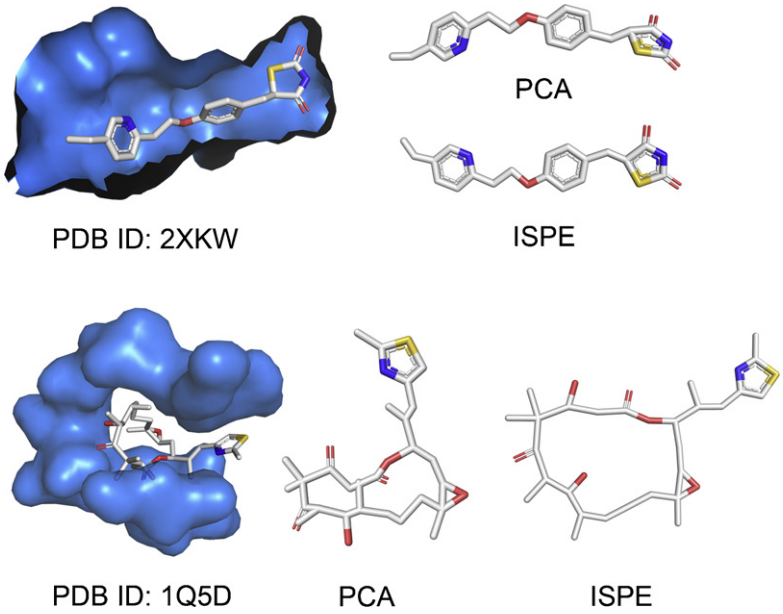

Fig. 6. Generation of 2D projections (right) for 3D molecular conformations (left). The top panel shows the results obtained by PCA and ISPE ($r_c = 0.2$) for the example of PPAR-gamma agonist pioglitazone bound to the receptor (PDB-ID [105] 2xkw [106]). The bottom panel provides the projections computed for epothilone D bound to cytochrome P450epoK (PDB-ID 1q5d [107]).

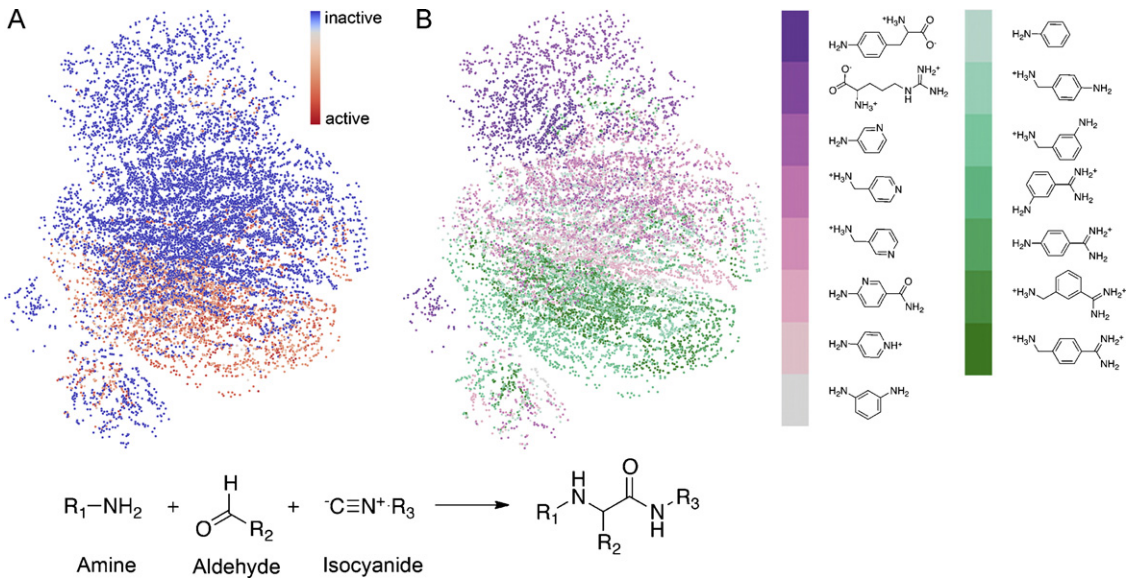

Fig. 7. 2D SNE projection of a combinatorial library containing 15,840 three-component Ugi reaction products (α-aminoacyl amide derivatives) formed from the condensation of an amine, aldehyde and isocyanide. In (A) the compounds' measured inhibitory activity against tryptase is shown by color-coding each compound (dot) from blue (inactive) to red (active). In (B) compound clusters containing the same amine building block (R_1) are highlighted in a different color. Data courtesy of Dr. Lutz Weber (Morphochem AG). Compound structures were standardized using the "wash" method in MOE v2010.10 and explicit hydrogen atoms were removed prior to descriptor calculation. Topological CATS descriptors were computed using the *speedcats* software (0–9 bonds, type-sensitive scaling) [108]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

also help to graphically investigate preliminary SARs. In Fig. 7B, the same ISPE projection is colored according to the primary amine identity used in the multi-component Ugi reaction. It becomes evident from looking at the color distribution that certain positively charged benzamidine derivatives seem to be preferred for blocking the serine protease tryptase, which has been long known from numerous medicinal chemistry projects [86].

7. Outlook

Maps of chemical space have demonstrated their usefulness for analyzing the diversity and complementarity of compound libraries, with particular emphasis on combinatorial compound collections and QSAR modeling [87–89]. Efforts to develop open platform for QSAR model generation and data analysis are ongoing [90,91], with visualization techniques playing an important role in activity prediction, extraction of ligand–target networks, structural diversity analysis, and cluster visualization. Structure–activity landscapes have received much attention recently [92,93], mainly driven by innovative visualization methods that allow for online monitoring of dynamic landscapes and fast and efficient embedding of chemical structures and picturing response surfaces [94,95]. Such methods, including our own visualization tool LiSARD (Ligand Structure Activity Relationship Display) [96], might become a valuable addition to the drug designer's toolbox. Latest developments include a study by Soto et al. who compared several mapping algorithms and suggest Correlative Matrix Mapping (CMM) as a potential method of choice for target-driven subspace mapping [97]. We are also witnessing continuing amalgamation of methods. For example, *Neighbor Embedding* XOM (NE-XOM), an extension to the *Exploratory Observation Machine* [98], is based on minimizing the Kullback–Leibler divergence of neighborhood functions in data and embedding space [99]. In this it is comparable to SNE combined with principles first encountered in SOM modeling. Numerous related concepts are being developed mainly in the context of machine-learning applications. We expect such innovative data mapping approaches to be studied for their transferability and practical usefulness in molecular modeling and drug discovery, thereby complementing automated virtual screening protocols for rapid focused library design and compound prioritization [100,101].

8. Further information/web resources

Open access/open source visualization tools

- R statistical computing: http://www.r-project.org/
- ParaView: http://www.paraview.org/
- Chembench [88]: http://chembench.mml.unc.edu
- ChemMine tools [91]: http://chemmine.ucr.edu
- iPHACE [102]: http://cgl.imim.es/iphace/
- ESOM [103]: http://databionic-esom.sourceforge.net/
- ChemSpaceShuttle [50]: http://gecco.org.chemie.uni-frankfurt.de
- CheS-Mapper: http://opentox.informatik.uni-freiburg.de/ches-mapper/

Chemical compound databases

- ChEMBL: https://www.ebi.ac.uk/chembl/
- PubChem: http://pubchem.ncbi.nlm.nih.gov/
- ChemBank: http://chembank.broad.harvard.edu/
- ChEBI: http://www.ebi.ac.uk/chebi/
- ChemDB: http://cdb.ics.uci.edu/
- ZINC [104]: http://zinc.docking.org/

Acknowledgments

The authors are grateful to Dr. Jan Hiss for stimulating discussion, and Dr. Petra Schneider and Dr. Lutz Weber for providing data sets. The Chemical Computing Group Inc. (Montreal, Canada) provided a research license of MOE software. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, FOR1406TP4).

References

[1] D.M. Maniyar, I.T. Nabney, B.S. Williams, A. Sewing, Data visualization during the early stages of drug discovery, J. Chem. Inf. Model. 46 (2006) 1806–1818.
[2] T.J. Howe, G. Mahieu, P. Marichal, T. Tabruyn, P. Vugts, Data reduction and representation in drug discovery, Drug Discov. Today 12 (2007) 45–53.
[3] B. Bienfait, J. Gasteiger, Checking the projection display of multivariate data with colored graphs, J. Mol. Graph. Model. 15 (1997) 203–215, 254–258.
[4] Y.A. Ivanenkov, N.P. Savchuk, S. Ekins, K.V. Balakin, Computational mapping tools for drug discovery, Drug Discov. Today 14 (2009) 767–775.
[5] J.L. Medina-Franco, K. Martinez-Mayorga, M.A. Giulianotti, R.A. Houghten, C. Pinilla, Visualization of the chemical space in drug discovery, Curr. Comput. Aided Drug Des. 4 (2008) 322–333.
[6] D.K. Agrafiotis, H. Xu, A geodesic framework for analyzing molecular similarities, J. Chem. Inform. Comput. Sci. 43 (2003) 475–484.
[7] M. Rupp, P. Schneider, G. Schneider, Distance phenomena in high-dimensional chemical descriptor spaces: consequences for similarity-based approaches, J. Comput. Chem. 30 (2009) 2285–2296.
[8] I.T. Jolliffe, Principle Component Analysis, Springer, New York, 2002.
[9] A. Linusson, M. Elofsson, I.E. Andersson, M.K. Dahlgren, Statistical molecular design of balanced compound libraries for QSAR modeling, Curr. Med. Chem. 17 (2010) 2001–2016.
[10] T.I. Oprea, J. Gottfries, Chemography: the art of navigating in chemical space, J. Comb. Chem. 3 (2001) 157–166.
[11] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107.
[12] Q. Li, T. Cheng, Y. Wang, S.H. Bryant, PubChem as a public resource for drug discovery, Drug Discov. Today 15 (2010) 1052–1057.
[13] K.P. Seiler, G.A. George, M.P. Happ, N.E. Bodycombe, H.A. Carrinski, S. Norton, S. Brudz, J.P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N.J. Tolliday, S.L. Schreiber, P.A. Clemons, ChemBank: a small-molecule screening and cheminformatics resource database, Nucleic Acids Res. 36 (2008) D351–D359.
[14] P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, C. Steinbeck, Chemical Entities of Biological Interest: an update, Nucleic Acids Res. 38 (2010) D249–D254.
[15] J. Chen, S.J. Swamidass, Y. Dou, J. Bruand, P. Baldi, ChemDB: a public database of small molecules and related chemoinformatics resources, Bioinformatics 21 (2005) 4133–4139.
[16] R. Gozalbes, A. Pineda-Lucena, Small molecule databases and chemical descriptors useful in chemoinformatics: an overview, Comb. Chem. High Throughput Screen. 14 (2011) 458–548.
[17] L.J. Bellis, R. Akhtar, B. Al-Lazikani, F. Atkinson, A.P. Bento, J. Chambers, M. Davies, A. Gaulton, A. Hersey, K. Ikeda, F.A. Krüger, Y. Light, S. McGlinchey, R. Santos, B. Stauch, J.P. Overington, Collation and data-mining of literature bioactivity data for drug discovery, Biochem. Soc. Trans. 39 (2011) 1365–1370.
[18] Y.C. Martin, J.L. Kofron, L.M. Traphagen, Do structurally similar molecules have similar biological activity? J. Med. Chem. 45 (2002) 4350–4358.
[19] A.M. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, 1990.
[20] F. Barbosa, D. Horvath, Molecular similarity and property similarity, Curr. Top. Med. Chem. 4 (2004) 589–600.
[21] T.I. Oprea, Chemical space navigation in lead discovery, Curr. Opin. Chem. Biol. 6 (2002) 384–389.
[22] R.E. Bellman, Adaptive Control Processes, Princeton University, Princeton, 1961.
[23] O.F. Güner, History and evolution of the pharmacophore concept in computer-aided drug design, Curr. Top. Med. Chem. 2 (2002) 1321–1332.
[24] P. Willett, Similarity searching using 2D structural fingerprints, Methods Mol. Biol. 672 (2011) 133–158.
[25] G. Schneider, S.S. So, Adaptive Systems in Drug Design, Landes Bioscience, Georgetown, 2001.
[26] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[27] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[28] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[29] H. Choi, S. Choi, Kernel isomap, Electron. Lett. 40 (2004) 1612–1613.
[30] M.A. Hibbs, N.C. Dirksen, K. Li, O.G. Troyanskaya, Visualization methods for statistical analysis of microarray clusters, BMC Bioinform. 6 (2005) 115.
[31] M.H. Law, A.K. Jain, Incremental nonlinear dimensionality reduction by manifold learning, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 377–391.
[32] G. Lee, C. Rodriguez, A. Madabhushi, Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies, IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008) 368–384.
[33] B.W. Higgs, J. Weller, J.L. Solka, Spectral embedding finds meaningful (relevant) structure in image and microarray data, BMC Bioinform. 7 (2006) 74.
[34] J. Ham, D.D. Lee, S. Mika, B. Schölkopf, A kernel view of the dimensionality reduction of manifolds, in: Proceedings of International Conference on Machine Learning, Banff, Canada, 2004, pp. 369–376.
[35] B. Schölkopf, A.J. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[36] Y. Sakiyama, The use of machine learning and nonlinear statistical tools for ADME prediction, Expert Opin. Drug Metab. Toxicol. 5 (2009) 149–169.
[37] O. Obrezanova, G. Csanyi, J.M. Gola, M.D. Segall, Gaussian processes: a method for automatic QSAR modeling of ADME properties, J. Chem. Inf. Model. 47 (2007) 1847–1857.
[38] M. Rupp, T. Schroeter, R. Steri, H. Zettl, E. Proschak, K. Hansen, O. Rau, O. Schwarz, L. Müller-Kuhrt, M. Schubert-Zsilavecz, K.R. Müller, G. Schneider, From machine learning to natural product derivatives that selectively activate transcription factor PPARgamma, ChemMedChem 5 (2010) 191–194.
[39] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, G. Schneider, Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors, J. Med. Chem. 48 (2005) 6997–7004.
[40] K. Hansen, D. Baehrens, T. Schroeter, M. Rupp, K.R. Müller, Visual interpretation of kernel-based prediction models, Mol. Inform. 30 (2011) 817–826.
[41] L. Rosenbaum, G. Hinselmann, A. Jahn, A. Zell, Interpreting linear support vector machine models with heat map molecule coloring, J. Cheminform. 3 (2011) 11.
[42] G. Schneider, P. Wrede, Artificial neural networks for computer-based molecular design, Prog. Biophys. Mol. Biol. 70 (1998) 175–222.
[43] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design – An Introduction, 2nd ed., Wiley-VCH, Weinheim, 1999.
[44] D.J. Livingstone, D.T. Manallack, I.V. Tetko, Data modelling with neural networks: advantages and limitations, J. Comput. Aided Mol. Des. 11 (1997) 135–142.
[45] D.J. Livingstone, Multivariate data display using neural networks, in: J. Devillers (Ed.), Neural Networks in QSAR and Drug Design, Academic Press, London, 1996, pp. 157–176.
[46] D.J. Livingstone, G. Hesketh, D. Clayworth, Novel method for the display of multivariate data using neural networks, J. Mol. Graph. 9 (1991) 115–118.
[47] G. Reibnegger, G. Werner-Felmayer, H. Wachter, A note on the low-dimensional display of multivariate data using neural networks, J. Mol. Graph. 11 (1993) 129–133.
[48] N. Brown, R.A. Lewis, Exploiting QSAR methods in lead optimization, Curr. Opin. Drug Discov. Develop. 9 (2006) 419–424.
[49] D.P. Visco Jr., R.S. Pophale, M.D. Rintoul, J.L. Faulon, Developing a methodology for an inverse quantitative structure–activity relationship using the signature molecular descriptor, J. Mol. Graph. Model. 20 (2002) 429–438.
[50] A. Gevehchi, A. Dietrich, P. Wrede, G. Schneider, ChemSpaceShuttle: a tool for data mining in drug discovery by classification, projection, and 3D visualization, QSAR Comb. Sci. 22 (2003) 549–559.
[51] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set descriptor space: a review, Altern. Lab. Anim. 33 (2005) 445–459.
[52] A. Tropsha, A. Golbraikh, Predictive QSAR modeling workflow, model applicability domains, and virtual screening, Curr. Pharm. Des. 13 (2007) 3494–34504.
[53] J.L. Melville, E.K. Burke, J.D. Hirst, Machine learning in virtual screening, Comb. Chem. High Throughput Screen. 12 (2009) 332–343.
[54] A. Schwaighofer, T. Schroeter, S. Mika, G. Blanchard, How wrong can we get? A review of machine learning approaches and error bars, Comb. Chem. High Throughput Screen. 12 (2009) 453–468.
[55] T.S. Schroeter, A. Schwaighofer, S. Mika, A. Ter Laak, D. Suelzle, U. Ganzer, N. Heinrich, K.R. Müller, Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules, J. Comput. Aided Mol. Des. 21 (2007) 485–498.
[56] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybern. 43 (1982) 59–69.
[57] D.B. Kirew, J.R. Chretien, P. Bernard, F. Ros, Application of Kohonen neural networks in classification of biologically active compounds, SAR QSAR Environ. Res. 8 (1998) 93–107.
[58] P. Schneider, Y. Tanrikulu, G. Schneider, Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing, Curr. Med. Chem. 16 (2009) 258–266.
[59] D.O. Hebb, The Organization of Behavior, Wiley & Sons, New York, 1949.
[60] J. Hertz, A. Krogh, R.G. Palmer, Introduction to the Theory of Neural Computation, Addison-Wesley, Redwood City, CA, 1991.
[61] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley & Sons, New York, 2001.
[62] T. Kohonen, Self-organization and Associative Memory, Springer-Verlag, Heidelberg, 1984.
[63] P. Schneider, K. Stutz, L. Kasper, S. Haller, M. Reutlinger, F. Reisen, T. Geppert, G. Schneider, Target profile prediction and practical evaluation of a Biginelli-type dihydropyrimidine compound library, Pharmaceuticals 4 (2011) 1236–1247.
[64] G. Schneider, W. Neidhart, T. Giller, G. Schmid, 'Scaffold-hopping' by topological pharmacophore search: a contribution to virtual screening, Angew. Chem. Int. Ed. Engl. 38 (1999) 2894–2896.
[65] A. Schüller, G. Schneider, Identification of hits and lead structure candidates with limited resources by adaptive optimization, J. Chem. Inf. Model. 48 (2008) 1473–1491.
[66] G. Schneider, P. Schneider, Navigation in chemical space: ligand-based design of focused compound libraries, in: H. Kubinyi, G. Müller (Eds.), Chemogenomics in Drug Discovery, Wiley-VCH, Weinheim, 2004, pp. 341–376.
[67] A. Yan, Application of self-organizing maps in compounds pattern recognition and combinatorial library design, Comb. Chem. High Throughput Screen. 9 (2006) 473–480.

[68] D. Digles, G.F. Ecker, Self-organizing maps for in silico screening and data visualization, Mol. Inf. 30 (2011) 838–846.
[69] J.D. Holliday, E. Kanoulas, N. Malim, P. Willett, Multiple search methods for similarity-based virtual screening: Analysis of search overlap and precision, J. Cheminform. 3 (2011) 29.
[70] A. Ultsch, Maps for the visualization of high dimensional data spaces, in: Proc. WSOM'03, Japan, 2003, pp. 225–230.
[71] P. Selzer, P. Ertl, Applications of self-organizing neural networks in virtual screening and diversity selection, J. Chem. Inf. Model. 46 (2006) 2319–2323.
[72] Z. Wu, G.G. Yen, A SOM projection technique with the growing structure for visualizing high-dimensional data, Int. J. Neural Syst. 13 (2003) 353–365.
[73] T. Furukawa, SOM of SOMs, Neural Netw. 22 (2009) 463–478.
[74] I.V. Tetko, Associative neural network, Methods Mol. Biol. 458 (2008) 185–202.
[75] S. Gupta, S. Matthew, P.M. Abreu, J. Aires-de-Sousa, QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties, Bioorg. Med. Chem. 14 (2006) 1199–1206.
[76] D.K. Agrafiotis, Stochastic proximity embedding, J. Comput. Chem. 24 (2003) 1215–1221.
[77] D.K. Agrafiotis, H. Xu, F. Zhu, D. Bandyopadhyay, P. Liu, Stochastic proximity embedding: methods and applications, Mol. Inf. 29 (2010) 758–770.
[78] M. Schmuker, G. Schneider, Processing and classification of chemical data inspired by insect olfaction, Proc. Natl. Acad. Sci. USA 104 (2007) 20285–20289.
[79] G. Hinton, S. Roweis, Stochastic neighbor embedding, Adv. Neural Inform. Process. Syst. 15 (2002) 833–840.
[80] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[81] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T.L. Griffiths, J.B. Tenenbaum, Parametric embedding for class visualization, Neural Comput. 19 (2007) 2536–2556.
[82] R. Memisevic, G. Hinton, Multiple relational embedding, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 913–920.
[83] B. Xie, Y. Mu, K. Huang, D. Tao, m-SNE: multiview stochastic neighbor embedding, IEEE Trans. Syst. Man Cybern. Part B 41 (2011) 1088–1096.
[84] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, J. Mach. Learn. Res. 11 (2010) 451–490.
[85] N. Kijoeng, J. Hongmo, C. Seungjin, Fast stochastic neighbor embedding: a trust-region algorithm, in: Proc. Intern. Joint Conf. Neural Networks (IJCNN), Budapest, Hungary, 2004, pp. 123–128.
[86] J. Stürzebecher, H. Vieweg, P. Wikström, D. Turk, W. Bode, Interactions of thrombin with benzamidine-based inhibitors, Biol. Chem. Hoppe Seyler 373 (1992) 491–496.
[87] L.B. Akella, D. DeCaprio, Cheminformatics approaches to analyze diversity in compound screening libraries, Curr. Opin. Chem. Biol. 14 (2010) 325–330.
[88] T. Walker, C.M. Grulke, D. Pozefsky, A. Tropsha, Chembench: a cheminformatics workbench, Bioinformatics 26 (2010) 3000–3001.
[89] N. Singh, R. Guha, M.A. Giulianotti, C. Pinilla, R.A. Houghten, J.L. Medina-Franco, Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository, J. Chem. Inf. Model. 49 (2009) 1010–1024.
[90] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I.I. Baskin, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I.V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, J. Comput. Aided Mol. Des. 25 (2011) 533–554.
[91] T.W. Backman, Y. Cao, T. Girke, ChemMine tools: an online service for analyzing and clustering small molecules, Nucleic Acids Res. 39 (2011) W486–W491.
[92] R. Guha, J.H. Van Drie, Structure–activity landscape index: identifying and quantifying activity cliffs, J. Chem. Inf. Model. 48 (2008) 646–658.
[93] R. Guha, The ups and downs of structure–activity landscapes, Methods Mol. Biol. 672 (2011) 101–117.
[94] P. Iyer, Y. Hu, J. Bajorath, SAR monitoring of evolving compound data sets using activity landscapes, J. Chem. Inf. Model. 51 (2011) 532–540.
[95] L. Peltason, P. Iyer, J. Bajorath, Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs, J. Chem. Inf. Model. 50 (2010) 1021–1033.
[96] M. Reutlinger, W. Guba, R.E. Martin, A.I. Alanine, T. Hoffmann, A. Klenner, J.A. Hiss, P. Schneider, G. Schneider, Neighborhood-preserving visualization of adaptive structure–activity landscapes and application to drug discovery, Angew. Chem. Int. Ed. 50 (2011) 11633–11636.
[97] A.J. Soto, G.E. Vazquez, M. Strickert, I. Ponzoni, Target-driven subspace mapping methods and their applicability domain estimation, Mol. Inf. 30 (2011) 779–789.
[98] A. Wismüller, The exploration machine: a novel method for analyzing high dimensional data in computer-aided diagnosis, Proc. SPIE 7260 (2009), 72600G.
[99] K. Bunte, B. Hammer, T. Villmann, M. Biehl, A. Wismüller, Neighbor embedding XOM for dimension reduction and visualization, Neurocomputing 74 (2011) 1340–1350.
[100] J.J. Irwin, Using ZINC to acquire a virtual screening library, Curr. Protoc. Bioinform. (2008) (Chapter 14: Unit 14.6).
[101] S.J. Campbell, A. Gaulton, J. Marshall, D. Bichko, S. Martin, C. Brouwer, L. Harland, Visualizing the drug target landscape, Drug Discov. Today 15 (2010) 3–15.
[102] R. Garcia-Serna, O. Ursu, T.I. Oprea, J. Mestres, iPHACE: integrative navigation in pharmacological space, Bioinformatics 26 (2010) 985–986.
[103] A. Ultsch, F. Moerchen, ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM, Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46, 2005.
[104] J.J. Irwin, B.K. Shoichet, ZINC – a free database of commercially available compounds for virtual screening, J. Chem. Inf. Model. 45 (2005) 177–182.
[105] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[106] J.J. Mueller, M. Schupp, T. Unger, U. Kintscher, U. Heinemann, Binding diversity of pioglitazone by peroxisome proliferator-activated receptor-gamma. Downloaded from: http://www.pdb.org, in press.
[107] S. Nagano, H. Li, H. Shimizu, C. Nishida, H. Ogura, P.R. Ortiz de Montellano, T.L. Poulos, Crystal structures of epothilone D-bound, epothilone B-bound, and substrate-free forms of cytochrome P450epoK, J. Biol. Chem. 278 (2003) 44886–44893.
[108] U. Fechner, G. Schneider, Optimization of a pharmacophore-based correlation vector descriptor for similarity searching, QSAR Comb. Sci. 23 (2004) 19–22.