Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Supervised extensions of chemography approaches: case studies of chemical liabilities assessment

Svetlana I Ovchinnikova[1,2], Arseniy A Bykov[1,2], Aslan Yu Tsivadze[1], Evgeny P Dyachkov[3] and Natalia V Kireeva[1,2*]

Abstract

Chemical liabilities, such as adverse effects and toxicity, play a significant role in modern drug discovery process. *In silico* assessment of chemical liabilities is an important step aimed to reduce costs and animal testing by complementing or replacing *in vitro* and *in vivo* experiments. Herein, we propose an approach combining several classification and chemography methods to be able to predict chemical liabilities and to interpret obtained results in the context of impact of structural changes of compounds on their pharmacological profile. To our knowledge for the first time, the supervised extension of Generative Topographic Mapping is proposed as an effective new chemography method. New approach for mapping new data using supervised Isomap without re-building models from the scratch has been proposed. Two approaches for estimation of model's applicability domain are used in our study to our knowledge for the first time in chemoinformatics. The structural alerts responsible for the negative characteristics of pharmacological profile of chemical compounds has been found as a result of model interpretation.

Keywords: Cheminformatics, Chemography, Applicability domain, Generative topographic mapping, Dimensionality reduction, Supervised generative topographic mapping, Isomap, Supervised Isomap

Background

During the past decade, computational technologies and predictive tools have been deeply integrated in the modern drug discovery process and changed this process extracting the useful knowledge embedded in the complex arrays of chemical and biological information to select the most promising compounds as early as possible and to reveal chemical liabilities in order to reduce the risk of late stage attrition [1,2]. Chemical liabilities, such as adverse effects and toxicity, play a significant role in modern drug discovery process. Methods to avoid or reduce chemical liabilities are an important target for drug discovery and development. Herein, we propose an approach combining several classification and chemography [3] methods to assess chemical liabilities *in silico* and to interpret obtained results in the context of impact of structural changes of compounds on their pharmacological profile. Model development has been performed in six different descriptor spaces for mutagenicity, carcinogenicity, acute toxicity and phospholipidosis data sets. A set of machine learning methods has been involved in model development encompassing well-known approaches with new ones. The combination of classification and data visualization is a key point for mechanistic model interpretation which allows one to understand which changes of the existing structures are required to improve target properties, to generate new hypothesis and, finally, to optimize the chemical structures. Over the years, a number of dimensionality reduction approaches [4-11] have been proposed and used in cheminformatics. The most known and widely used among these methods are Principal Component Analysis [12], Multidimensional Scaling (MDS) [13,14], Self-Organizing Maps (SOM) [15], Stochastic Proximity Embedding [16-18], Stochastic Neighbor Embedding [19,20], Sammon Mapping [21] and Generative Topographic Mapping (GTM) [22-24]. In this study, Generative Topographic Mapping and Isomap as well as their supervised extensions have been involved. Recently, the unsupervised

* Correspondence: nkireeva@gmail.com
[1]Frumkin Institute of Physical Chemistry and Electrochemistry RAS, Leninsky pr-t 31-4, 119071 Moscow, Russia
[2]Moscow Institute of Physics and Technology, Institutsky per., 9, 141700 Dolgoprudny, Russia
Full list of author information is available at the end of the article

implementations of these approaches have been used in a number of studies in chemoinformatics [25-32]. These two representatives of nonlinear dimensionality reduction methods are related to two different families: distance-based approaches and topology based approaches. Isomap reduces the dimensionality of data by using distance preservation as the criterion, that is intuitively understandable and easy to compute. GTM is related to the topology based techniques. This group of methods tries to preserve topology principle that is concerned to relative proximities: compounds which are close in the data space remain close in the data visualization model. Topology preservation usually is considered as more powerful and in the same time more complex comparing with distance preservation [4]. The comparison of used techniques on the considered data is performed in this study. Support vector machines (SVM) [33], GTM and probabilistic neural networks (PNN) [34] have been used for the development of classification models. Two applicability domain of models' approaches (AD) are involved in our study in order to assesses the model's limitation in prediction of new data in order to reliably predict those data that are structurally similar to the training set compounds used for model development. Recently, several different AD approaches have been proposed [35-49]. Here, we use the representatives of two families of AD methods: distance-based (Ball) [50] and probability-based (Local Outlier Factor LOF) [51].

Here, to our knowledge for the first time, we propose supervised extension of Generative Topographic Maps [52] that can be used as a universal tool to visualize the chemical space and to develop classification models. New approach for projecting new data using supervised Isomap [53] without re-building models from the scratch has been developed. The evaluation of the performance of the dimensionality reduction techniques and introduced descriptor spaces to separate different activity classes has been monitored by three parameters, two of them have been used in cheminformatics for the first time.

Materials and methods

Data preparation

Data preparation has been carried out using recommendations published in [54]. Chemaxon Standardizer [55] and Instant JChem [56] software have been used for the data preparation. Using Standardizer, the explicit hydrogen atoms have been removed, the structures have been aromatized and neutralized. Four data sets have been used in our study.

Mutagenicity

Ames mutagenicity data from a study by Kazius et al. [57]. The data set contained 2367 active and 1888 inactive compounds. External test set consists of 1164 active and 2167 inactive compounds.

Carcinogenicity

Data was collected from the distributed ISSCAN Database (part of structure-searchable toxicity DSSTox public database network [58]). The database has been specifically designed as an expert decision support tool and includes the carcinogenicity classification "calls" to guide the application of SAR approaches. Collected data set encompass 1088 chemical structures containing 648 compounds annotated as actives and 440 as inactive compounds. External test set [25] contains 359 actives and 141 inactives.

Phospholipidosis

A set of 100 phospholipidosis-inducing compounds and 82 negative drug-like compounds were taken from [59], where the active compounds have been observed to act on a range of species (humans, rats, mice, dogs, rabbits, hamsters and monkeys) and on a variety of tissue types (lungs, kidney and liver). External test set from [60] contains 141 active and 359 inactive compounds.

Acute toxicity

Data from EPA Fathead Minnow Acute Toxicity Database [61] after data preparation stage containing 612 compounds (578 actives and 34 inactives). This database was generated by the U.S. EPA Mid-Continental Ecology Division (MED) for the purpose of developing an expert system to predict acute toxicity from chemical structures based on mode of action considerations.

Descriptors

In this study, six descriptor types have been involved in model development. ISIDA package [62] has been represented by two different descriptor types: ***(i)*** ISIDA Property-Labeled Fragment Descriptors (IPLF) [63] (atom-centered fragments (augmented atoms) of radius 1 to 3 colored by pH-dependent pharmacophores and ***(ii)*** subclass of ISIDA Substructural Molecular Fragments (SMF) [62] consisting of the shortest topological paths with explicit representation of only terminal atoms and bonds, where the values of minimal n_{min} and maximal n_{max} number of atoms varied from 2 to 15. 2D descriptors of Molecular Operating Environment (MOE 2D) [64] containing different physical properties, subdivided surface areas, atom and bond counts, Kier&Hall connectivity and Kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors and partial charge descriptors were involved in model development. The CDK (Chemistry Development Kit) MACCS keys and extended fingerprints (EF) were computed using the RCDK package [65] of the R software [66]. Finally, Dragon software [67] has been used for molecular descriptors calculations. Constant and nearly constant descriptors were removed. Detailed table

with the final number of descriptors for each data set and descriptor type is represented in supporting information.

Methods

Classification methods

Support Vector Machines (SVM)

SVM [68,69] is a supervised learning method commonly used for classification and regression and based on statistical learning theory of Vapnik–Chervonenkis [70,71]. Projecting the original data described by means of descriptor vectors to a higher dimensional feature space SVM achieves distinct separation between considered classes of compounds finding the optimal position of the separating hyperplane between the instances from the classes.

Generative Topographic Mapping (GTM)

GTM is a specific unsupervised density network based on generative modeling. It can be considered as probabilistic extension of Kohonen Self-Organizing Maps. Like SOM, it operates with a grid of K nodes, which can be considered as analogs of nodes in SOM. GTM creates a generative probabilistic model in the high-dimensional data space RD by placing a radially symmetric Gaussian with zero mean and inverse variance β around projections of the latent space centers which approximating the data density. The nonlinear GTM transformation from the latent space to the data space is defined using a Radial Basis Function (RBF) network. Thus, each node is projected to the center of Gaussian belonging to the manifold (two-dimensional flexible sheet located in the high-dimensional space in such a way to cover the data points by stretching or compressing) embedded in the data space. This manifold can be considered as a representation of the latent space in the data space. The coordinates of the Gaussians are computed as a linear combination of Gaussian basis functions and for the point $\boldsymbol{x}$ in the latent space its projection to the data space can be defined as:

$$y = \mathrm{W}\boldsymbol{\phi}(\boldsymbol{x}) \tag{1}$$

where W- the output weights of RBF.

It relates the real data in the chemical space with manifold points. Thus, any point of the latent space $\mathbb{R}^L$ has its own projection in a data space $\mathbb{R}^D$ obtained by non-linear parameterized mapping $y(\boldsymbol{x}, \mathrm{W})$.

The mapping function $y(\boldsymbol{x}, \mathrm{W})$ is continuous, which leads to the topographic ordering of the projected points, i.e. two points that are close in the latent space are also close in the data space. Defining a probability distribution over the latent space induces the corresponding distribution over the manifold in the data space and, thus, imposes the probabilistic relationships between two spaces.

The iterative Expectation-Maximization algorithm (EM-algorithm) is used to find the parameters of RBF network (W and β) maximizing the, so called, log likelihood function which measures a correspondence between the data distribution and the model.

$$\mathcal{L}(\mathrm{W}, \beta) = \sum_{n=1}^{N} ln\left\{\frac{1}{K}\sum_{i=1}^{K} p(\boldsymbol{t_n}|\boldsymbol{x_i}, \mathrm{W}, \beta)\right\} \tag{2}$$

where $\mathcal{L}$ - log likelihood function, β - inverse of variance, W - the output weights of RBF, K - number of the nodes, N - number of compounds, $p(\boldsymbol{t_n}|\boldsymbol{x_i}, \mathrm{W}, \beta)$ – prior probability generated in a point $\boldsymbol{t_n}$ in the data space by the Gaussian with a center in $y(\boldsymbol{x_i}, \mathrm{W})$.

Activity profile of a chemical compound can be assessed starting from the values of the class-conditioned probability distribution function $p(\boldsymbol{t}|C_k)$ computed for each class C_k, where $\boldsymbol{t}$ is its molecular descriptor vector. Such function can be build, for each activity class, by training a separate GTM model on the data belonging to class C_k. The class-conditioned probabilities $p(\boldsymbol{t}|C_k)$ can be used for computing posterior probabilities of class membership $P(C_k|\boldsymbol{t})$ for a given compound using the Bayes theorem:

$$P(C_k|\boldsymbol{t}) = \frac{p(\boldsymbol{t}|C_k)\cdot P(C_k)}{p(\boldsymbol{t})} \tag{3}$$

where $P(C_k) = \frac{N_k}{N_{tot}}$ is a prior probability of class membership (N_k – the number of compounds belonging to class C_k; N_{tot} – the total number of compounds), whereas $p(\boldsymbol{t})$, the marginal probability density function, is the normalization factor:

$$p(\boldsymbol{t}) = \sum_{k} p(\boldsymbol{t}|C_k)\cdot P(C_k) \tag{4}$$

The latter ensures that the estimated posterior probabilities are normalized. By applying Function 3 to each class C_k one can assess the posterior probability of class membership for each compound. According to statistical decision theory [72], the optimal class assignment is determined by the maximal value of posterior class probabilities $P(C_k|\boldsymbol{t})$.

Probabilistic Neural Networks (PNN)

PNN [34] belongs to a group of feed-forward neural network algorithms. It was derived from Bayesian Networks [73] and Kernel Discriminant Analysis [74]. PNN consists of four layers: input layer, pattern layer, summation layer and output layer.

An input layer represents the input vector, e.g. a compound from a test set. Each compound is attributed to a single neuron of pattern layer, for which its descriptors

represent a weight vector. Therefore all pattern neurons can be marked with the class labels of corresponding compounds. Input layer interconnected with a pattern layer, thus each pattern unit forms a dot product Z of an input vector and its weight vector. Z is propagated to the network activation function $e^{\frac{Z-1}{\sigma^2}}$ and the result is outputted to the summation layer. Each neuron in the summation layer is connected to pattern units of the corresponding class. This layer performs simple summation of the inputs from the pattern layer. The output layer is a two-input layer, which produces a binary output. It takes into account the contribution for each class of inputs. The output is a 1 (positive identification) for that class and a 0 (negative identification) for non-targeted classes. In fact, there's no training required since the compounds of the training set are considered as the weights to the hidden layer of the network. As no training required, classifying an input vector is fast, depending on the number of classes and compounds in use. PNNs have some advantages comparing with multilayer perceptron networks: they are faster, relatively insensitive to outliers and generate probability scores.

Dimensionality reduction methods

Supervised Generative Topographic Mapping (s-GTM)

GTM performs visualization by inversing mapping from the data space to the latent space (unbending this flexible sheet into the rectangular 2D map). For this Bayes theorem is used. Thus, for each molecule GTM calculates its probability to be located in the given point of this map represented by the latent space and visualizes this molecule according this probability.

In order to make manifolds location in the data space dependent on distribution not only of the whole data set, but also of each class, a new supervised training procedure was performed. Each iteration consists of two major steps.

On the first step latent points are ascribed to one of the data classes in consideration. To this end we calculate responsibilities r_{kn} (i.e. posterior probabilities that data point $\boldsymbol{t}_n$ was generated by the component $\boldsymbol{x}_k$).

$$r_{kn} = p(\boldsymbol{x_k}|\boldsymbol{t_n}, \mathrm{W}, \beta) = \frac{p(\boldsymbol{t_n}|\boldsymbol{x_k}, \mathrm{W}, \beta)p(\boldsymbol{x_k})}{\sum_{i=1}^{K} p(\boldsymbol{t_n}|\boldsymbol{x_i}, \mathrm{W}, \beta)p(\boldsymbol{x_i})} \quad (5)$$

where $p(\boldsymbol{x_k}) = \frac{1}{K}$.

For each latent point $\boldsymbol{x}_k$ the following sums are calculated

$$S_j = \frac{1}{N_j}\sum_{i=1}^{N_j} r_{ki}, j = 1, 2 \quad (6)$$

where index j refers to one of the classes, N_j – is a number of compounds in this class.

The latent point is associated with the class with the largest sum of responsibilities, only in a case when the difference between the sums is greater than the threshold value *thr*, which is an external parameter of the method. If not, the latent point remains unlabeled. To assure the formation of clusters of similarly labeled latent points, the influence of neighbor latent points is taken into account by decreasing the threshold value if the latent point on previous iteration had neighbors associated with the class which responsibility sum is larger on the current iteration and increasing it if the neighbors are from the opposite class.

The second step contains movement of the latent points projections towards data points of the corresponding class by adjusting the RBF network. The sum $\bar{T}$ of vector distances from the latent point $\boldsymbol{x}_k$ projection to all the data points $\boldsymbol{t}_n$, for which $r_{kn} > rr$, is calculated. Here, *rr* denotes the responsibility radius, another external parameter of s-GTM. If the data point belongs to the class opposite to that of the latent point, the corresponding distance is multiplied by -1 (thereby, the vector form the data point to the latent point is obtained). The desired new coordinates $\bar{P}'$ of the latent point projection are defined the following way:

$$\bar{P}' = \bar{P} + \frac{\bar{T}}{N} \quad (7)$$

Then RBF network is trained using the coordinates of $\boldsymbol{x}_k$ in the latent space as input and $\bar{P}'$ as a target.

Supervised GTM has a number of external parameters that have a great influence on the model development. Main parameter for latent points' colorization is the threshold value. It should be low enough, in order to allow a considerable amount of latent points to get labeled. The maximum value can be found from analyzing the responsibility matrix and strongly depends on the number of latent points: the larger is their number, the lower should be the threshold value.

The influence of the color of neighbor latent point is defined by additional compound for threshold calculation:

$$thr'_{1,l} = thr + \frac{N_2 - N_1}{\rho} \quad (8)$$

where *thr* is an original value, N_1 and N_2 – number of neighboring latent points of class 1 and 2 respectively, ρ - an external parameter, $thr'_{1,l}$ - is a threshold value, specific for class 1 and latent point l. This means, that latent point l will be labeled as class 1, if

$$S_1 > S_2 + thr'_{1,l} \quad (9)$$

It is obvious, that parameter ρ is required to bring both terms of Formula 9 to similar scale. Surprisingly, in quite a wide range it has small impact on the model, but

can be very useful for imbalanced data to prevent all the latent point to be marked by the same class label. It should be altered for fine optimization or in case if no similarly labeled clusters of latent points are formed during the training process.

S-Isomap

Isomap [75] is a low-dimensional embedding method. It implies that data are disposed along a manifold with a dimensionality d less than dimensionality d_o of the original data space. Our aim is to "unroll" the manifold into a d-dimensional space, so that data points, which are close to each other on the manifold remain close, and remote points – stay remote. To this end, we replace Euclidian distance with geodesic one – the length of the shortest curve between two points that lies on the manifold.

Isomap algorithm consists of three steps. On the first step we define k nearest neighbors of each compound and assume that Euclidian distances between them are small and, thus, are nearly equal to corresponding geodesic distances. This assumption allows us to create a weighted graph where only the vertices that are nearest neighbors are connected and the length of each edge equals the corresponding distance. This graph is not always connected and in this case the largest connected part is taken for the next step. After the graph has been constructed we compute shortest distances between its vertices. Then obtained distance matrix is used for multidimensional scaling (MDS) [13,14] from original to d-dimensional space. To minimize the cost function in MDS coordinates of compounds in the new space should be set to the top d eigenvectors of the matrix $\tau\left(\tilde{D}\right)$ [76], where $\tilde{D}$ is a matrix of pairwise distances between training points and τ is an operator, that converts distances to inner products. For visualization purpose we set $d = 2$.

Supervised extension of Isomap was proposed in [53]. It differs from the original algorithm in its first step. Instead of Euclidian distance $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ a new measurement of compounds' dissimilarity is calculated.

$$D(\boldsymbol{x}_i,\boldsymbol{x}_i) = \begin{cases} \sqrt{1-e^{\frac{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\beta}}}, if\ y_i = y_j \\ \sqrt{e^{\frac{d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\beta}}} - \alpha, if\ y_i \neq y_j \end{cases} \quad (10)$$

Here y_i denotes class label of compound $\boldsymbol{x}_i$, and β is a parameter that prevents $D(\boldsymbol{x}_i,\boldsymbol{x}_j)$ from increasing too fast. β should depend on data density and average Euclidian distance between all pairs of data points is usually used. The parameter α gives some chance to the points from different classes to be more close to each other.

After new distances have been calculated k-nearest neighbors are defined and weighted graph is constructed in the way it is done in non-supervised algorithm.

A way to extend nonsupervised Isomap to new points was proposed in [77,78]. There coordinates of new points are calculated as

$$e_k(\boldsymbol{x}) = \frac{1}{2\sqrt{\lambda_k}}\sum_i \upsilon_{ki}\left(E_{x'}\left(\tau\left(\tilde{D}\left(\boldsymbol{x}',\boldsymbol{x}_i\right)\right)\right) - \tau\left(\tilde{D}(\boldsymbol{x}_i,\boldsymbol{x})\right)\right) \quad (11)$$

where λ_k is eigenvalues and υ_{ki} - coordinates of the corresponding eigenvectors of the matrix $\tau\left(\tilde{D}\right)$, operator $E_{x'}$ denotes average over the data set. To make this work for S-Isomap we take into consideration Eq. 11, while computing $\tilde{D}(\boldsymbol{x}_i,\boldsymbol{x})$ – geodesic distance from and external point $\boldsymbol{x}$ to the training point $\boldsymbol{x}_i$. We assume that the distance from $\boldsymbol{x}$ to its $\boldsymbol{k}$ nearest neighbors of $\boldsymbol{x}$ is small enough to make not much difference between two parts of Eq. 11, and so we can use their average as a geodesic distance from $\boldsymbol{x}$ to its k nearest neighbors. Other geodesic distances are found from matrix $\tilde{D}$ by computing the shortest paths as it has been done while training the model. If value $\frac{d^2(\boldsymbol{x},\boldsymbol{x}_i)}{\beta}$ is too large (which happens when average distances between compounds in the original data space are much exceed one), additional coefficient β_1 can be used for both training the model and extending it to the new points. In this case the parameter β in Eq. 11 is replaced with $\beta_1\beta$.

Applicability domain approaches

Ball

Ball [50] is a distance-based method for outlier detection. It uses $\boldsymbol{L}^p$ –metric, in which distance between compounds $\boldsymbol{x}$ and $\boldsymbol{y}$ in feature is space denoted by Formula 11.

$$dist_{\boldsymbol{L}^p}(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_i |x_i - y_i|^p\right)^{1/p} \quad (12)$$

The algorithm optimizes the weight vector $\boldsymbol{w}$ the following way:

$$\begin{aligned} &\min\rho \\ &s.t.\sum_j w_j\left|x_{ij} - a_j\right|^p \le \rho \\ &\sum_j w_j = 1, w_j \ge 0 \end{aligned} \quad (13)$$

where $\boldsymbol{a}$ is a centroid of the data points and x_{ij} denotes the coordinate j of the compound $\boldsymbol{x}_i$.

After $\boldsymbol{w}$ is optimized, the compounds $\boldsymbol{x}_i$ for which $\sum_j w_j|x_{ij}-a_j|^p$ is the largest are considered as outliers. In other words this method fit $\boldsymbol{L}^p$ "ball" around the data. This "ball" separates targets from outliers. Figure 1a demonstrates the case of 2-dimensional feature space with $w_1 = w_2$.

Local Outlier Factor (LOF)

LOF is a probability based method for outlier detection in a multidimensional dataset [51]. It operates with local densities of objects in the dataset by using the definition of local reachability density and calculates value of "local outlier factor" that indicates the degree of object's dissimilarity to other compounds in the data set.

To define the local reachability density we should first introduce some other concepts. We call k-distance of the object p ($dist_k(p)$) the smallest value for which there are at least k objects besides p with a distance from p smaller or equal to $dist_k(p)$. K-distance neighborhood of an object p ($N_k(p)$) is a set of objects, not including p, whose distance from p does not exceed $dist_k(p)$. Let us specify that the cardinality of $N_k(p)$, which we also denote as $|N_k(p)|$, can be greater than k in case, when in $N_k(p)$ exist two or more objects whose distances from p are equal to $dist_k(p)$. Reachability distance of object p with respect to object o ($rdist_k(p,o)$) is the maximum value between k-distance of o and distance from o to p. The idea of reachability distance is illustrated in Figure 1b.

Local reachability density can be defined as

$$lrd_k(p) = \left(\frac{\sum_{o\in N_k(p)} rdist_k(p,o)}{|N_k(p)|}\right)^{-1} \quad (14)$$

The local outlier factor is an average of the ratios of the local reachability densities of objects to those of object's k nearest neighbors (Eq. 15).

$$LOF_k(p) = \frac{\sum_{o\in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \quad (15)$$

In [51] is shown, that LOF of objects that lie 'deep' inside a cluster approximately equals to 1. It is also shown that in majority of cases k can be chosen so that for all objects that belong to some cluster of objects LOF approximately equals to one, and for any other object it significantly differs from one. This fact allows us to detect compounds that do not belong to any cluster and so can be called outliers.

Experimental

The predictive performance of developed classification models was assessed using five-fold external cross-validation (5-CV) procedure considering Balanced Accuracy (BA) value [79] as a criterion of the predictive performance of the models. BA is an average of two other criteria, Sensitivity and Specificity, which were designed to assess model's ability to identify compounds from a certain class (active or positive for Sensitivity and inactive or negative for Specificity) disregarding its behavior for the other class. The combination of Sensitivity and Specificity should be able to compensate possible imbalance in the dataset.

$$BA = \frac{1}{2}(Sens + Spec) = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right) \quad (16)$$

where *Sens* is Sensitivity, *Spec* is Specificity, *tp* stands for *true positive* rate (e.g. the number of correctly predicted active compounds), *tn* – for *true negative* (correctly predicted inactive compounds), *fp* – for *false positive*

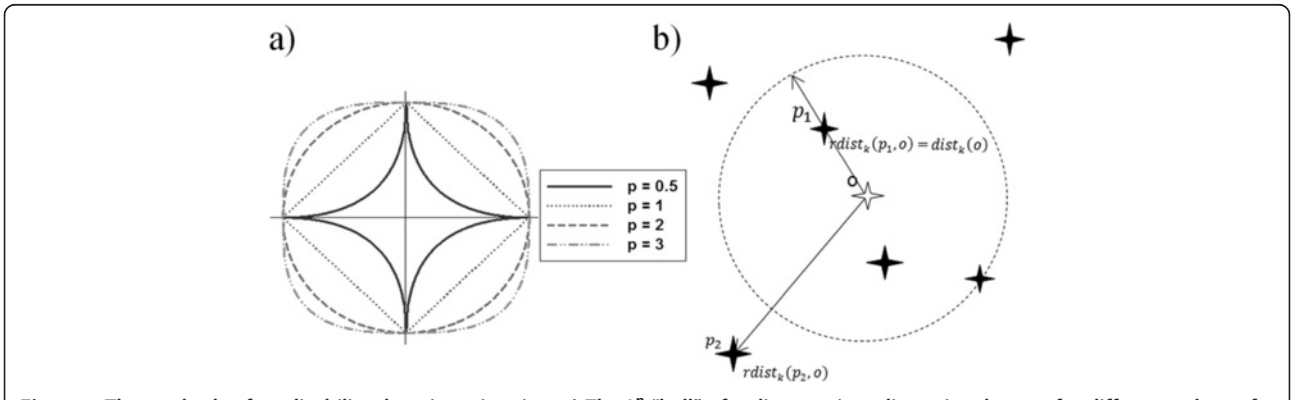

Figure 1 The methods of applicability domain estimation: a) The L^p "ball" of radius one in 2-dimentional space for different values of p and $w_1 = w_2$; b) the reachability distance of objects p_1 and p_2 with respect to object o for $k = 3$.

(inactive compounds that've been predicted to be active), *fn* – for *false negative* (active compounds that've been identified as inactive ones).

LibSVM [80] was used for developing SVM models, two its external parameters ν and γ were varied from 0.01 to 0.91 and from 2^{-11} to 2^{3} respectively.

GTM models were built with the help of the Netlab [81] package. This implementation can't work with large number of descriptors, so the Principal Component Analysis was introduced beforehand. Here, the following external parameters were gone over. Number of first principal components, that were retained, was varied from 20 to 60, number of latent points – from 5^2 to 50^2, number of radial basis network centers – from 2^2 to 7^2.

PNN was implemented in Classification Toolbox for use with MATLAB [82]. Its only external parameter Gaussian width was chosen from the range of [0; 1].

Among all the developed models for each combination of dataset, descriptor type and applied method one with the highest Balanced Accuracy was selected for further analysis.

For s-GTM the value of threshold was tried from 0 to 0.2, in most cases we used $\rho = 40$ or $\rho = 30$. The only external parameter in the movement step (responsibility radius, *rr*) has a great influence on the model. Too small values leads to small changes in the model compared to unsupervised GTM, too big – to mapping all compounds into a single point. This parameter was sorted out in a large range.

The performance of data visualization has been monitored with three quantitative measures. Each of them is normalized to vary from 0 to 1 and can be computed for a data set where the information about the classes is available.

Γ-score

Γ-score [26] takes into account k nearest neighbors of each projection. The more neighbors of each point belong to the same class, the higher is Γ-score. Thus, this score characterizes the ability of a model to produce similar-structure clustering in a visualization. To compute Γ-score one need to take the following steps. First, for each compound v_l $G(l, k)$ should be computed:

$$G(l,k) = \frac{1}{k}\sum_{j=1}^{k} g(v_l, j) \tag{17}$$

where k is the number of nearest neighbors, which is an external parameter, $g(v_l, j) = 1$ if the jth nearest neighbor of v_l in the visualization space belongs to the same class as v_l, g $(v_l j) = 0$ otherwise. Then for each class i γ_i (k) is defined as

$$\gamma_i(k) = \frac{1}{n_i}\sum_{l=1}^{n_i} G(l,k) \tag{18}$$

where n_i is a number of compounds of class i. And finally the Γ-score is

$$\Gamma(k) = \frac{1}{N}\sum_{i=1}^{N}\gamma_i(k) \tag{19}$$

where N is a number of classes.

Distance Consistency (DSC)

DSC [83] is based on the distances from points to the centroid of each class. It is higher when more points are closer to the centroid of the corresponding class, then to any other. The score is equal to 1, if the model projects compounds into separate clusters, one for each class. The computation of DSC is similar to the computation of Γ-score, but instead of $g(v_l, j)$ the centroid distance (CD) is used. Beforehand for each class i one need to find the coordinates of its centroid c_i. Then $CD(v_l, c_i) = 1$ if the closest to v_l is the centroid c_i and v_l belongs to class i and $CD(v_l, c_i) = 0$ otherwise. Then for each class i

$$C(i) = \frac{1}{n_i}\sum_{l=1}^{n_i} CD(v_l, c_i) \tag{20}$$

$$DSC = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{21}$$

Distribution Consistency (DC)

DC [83] estimates the overlapping of classes. It divides a map into separate areas and treats them independently. For each area the value of entropy is computed, which is 0, if all the points in the area share one class label, and reaches maximum, when every class is represented in the area by equal number of points. For DC computation the conception of entropy of the region R is to be introduced.

$$H_R = -\sum_{i=1}^{N}\frac{p_i}{\sum_i p_i}\log_2\left(\frac{p_i}{\sum_i p_i}\right) \tag{22}$$

Here, p_i is a number of molecules of class i in the region R. And the value of DC is defined the following way

$$DC = 1 - \frac{1}{Z}\sum_{R} p_R H_R \tag{23}$$

p_R is the whole number of molecules in the region R and a coefficient $Z = n\log_2 N$ is used to range DC from 0

to 1. In this work to obtain the required regions we divided the visualization map into 15×15 equal sized rectangles.

Results and discussion

Classification models performance

As one can see in Figure 2, for three out of four considered datasets, the best predictive performance was demonstrated by the Support Vector Machine approach (carcinogenicity – 68%, mutagenicity – 83%, phospholipidosis – 82%). Yet, in prediction of acute toxicity GTM significantly outperformed SVM (Balanced Accuracy reached 86% for GTM and 75% for SVM models). It is also seen that for GTM approach IPLF descriptors shown to be less effective than others, while applying molecular fingerprints for both SVM and GTM approaches led to high values of Balanced Accuracy. The behavior of accuracy for acute toxicity predictions significantly differs from those for other data sets. Molecular fingerprints here showed nearly the worst results among all the types of descriptors in this study (63% for SVM and 71% for GTM), while the best predictive performance was achieved using descriptors of the MOE and Dragon packages. The corroboration and a possible explanation of this fact may be given by found in the attempt of detection of structural alerts that is given further in this chapter. Implementation of MACCS descriptors failed to mark out any fragments that are responsible for toxic activity of compounds. The reason for this may be the imbalance of this particular data set. The deficiency of inactive compounds leads to difficulties in determining whether the presence of a fragment in several inactive compounds is an accident. Though in most cases SVM outperforms GTM, the analysis of its work is obstructed by the lack of intrinsic information about the predictive decisions. GTM, on the other hand, not only gives easily interpretable probability distribution for each compound, but also can be used as a tool for data visualization and outlier detection.

PNN may be considered a compromise between the lack of method's internal information of SVM and the decrease of accuracy of GTM. It is not such a universal tool as GTM but slightly outperforms it (up to 6% for mutagenicity). At the same time, PNN makes less

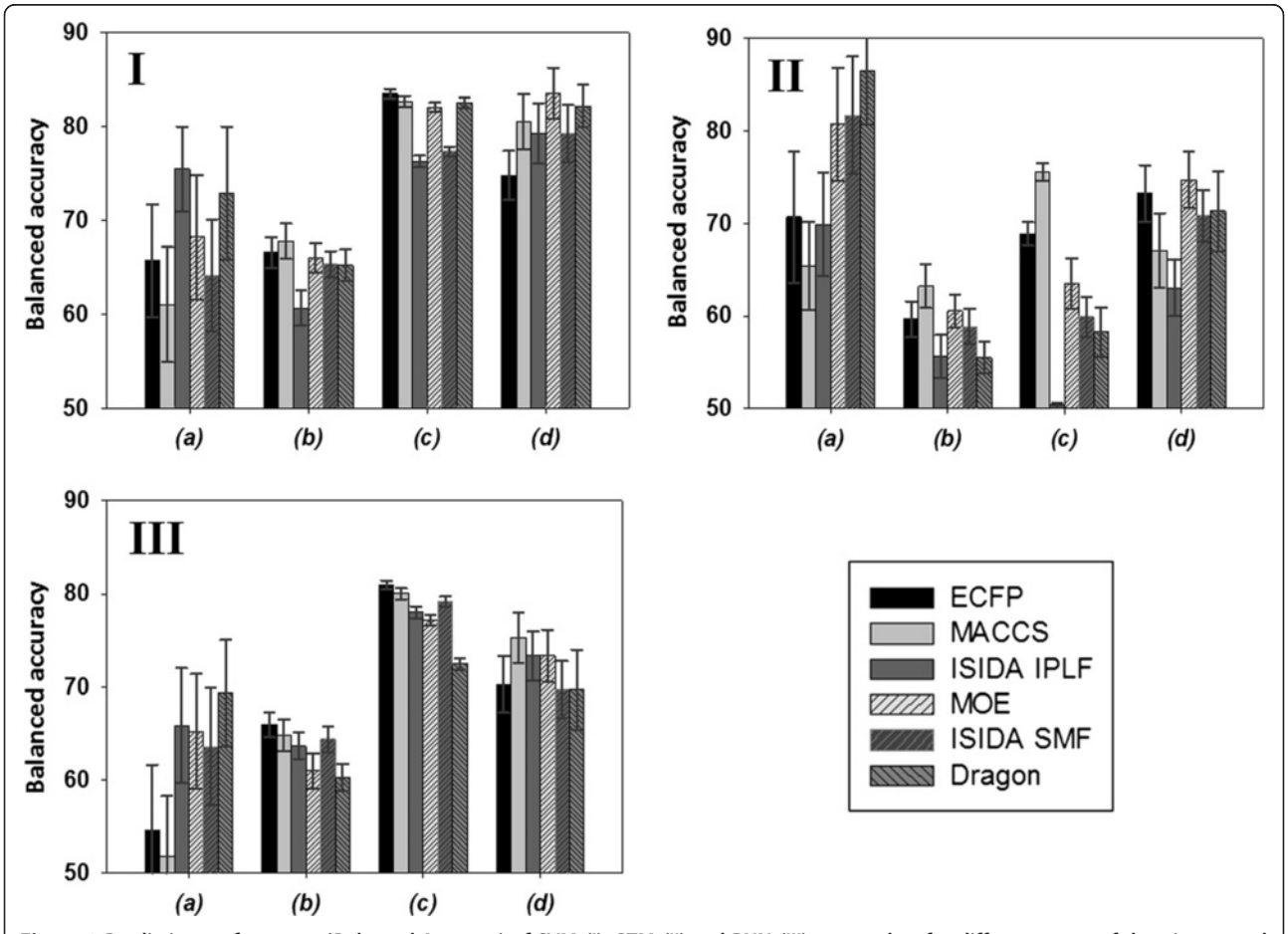

Figure 2 Predictive performance (*Balanced Accuracy*) of SVM *(I)*, GTM *(II)* and PNN *(III)* approaches for different types of descriptors and different datasets: acute toxicity *(a)*, carcinogenicity *(b)*, mutagenicity *(c)*, phospholipidisis *(d)*.

accurate predictions then SVM, but allows one to look through the background of each decision by analyzing pattern and decision layers. There is a similarity in behavior of SVM and PNN.

Dependence of Balanced Accuracy from datasets and descriptor types obtained by PNN is turned out to be similar to that of SVM, but not of GTM, though both PNN and GTM are neural networks.

The considered data sets were previously studied by other teams. Thus, classification of the acute toxicity data set has been performed in [84]. The compounds have been divided into classes differently than in our study and in the original database. A set of different machine learning approaches including several types of neural networks as well as SVM, Decision Trees and Gene Expression Programming have been applied for classification purposes. Corresponding Balanced Accuracy values of the developed models varied in the range from 0.85 to 0.93. A number of studies [85,86] with the regression analysis have been published including the original publication of this data set [61]. The carcinogenicity data involved in this study has been used in QSAR studies mostly as a source of further data retrieval. It has been used, for example, as a part of considered data in [87]. A thorough analysis of the mutagenicity data set including the applicability domain estimation has been performed in [40]. The direct comparison of the obtained results performance is straitened because of the difference in the statistical parameters used. Comparable results (obtained by combination of ECFP descriptors with Random Forest and Nearest Neighbor classifiers) have been recently reported in [88]. In [59] SVM and Random Forest were applied for phospholipidosis prediction. There Matthews Correlation Coefficient was used to assess the results performance, and its values varied up to 0.72 that outperforms the maximum value of this parameter in our study.

Predictions of models developed on IPLF, ECFP and SMF descriptors were analyzed. The numbers of compounds containing a certain descriptor d, predicted to be active n^d_{act} and inactive n^d_{inact}, were calculated for each descriptor. Then the corresponding fractions of compounds were calculated as

$$fr^d_{act} = \frac{n^d_{act}}{n_{act}}, fr^d_{inact} = \frac{n^d_{inact}}{n_{inact}} \quad (24)$$

where n_{act} and n_{inact} are total number of compounds, predicted to be active or, respectively, inactive, by the model at issue. The rare descriptors with $fr^d_{act} + fr^d_{inact} < 0.05$ were excluded from further consideration. Among other descriptors were selected for further analysis those with $\frac{fr_{act}}{fr_{inact}} > 2.5$. Fragment descriptors used for the prediction of compounds as actives by all three classification methods are demonstrated in Figure 3.

MACCS descriptors were not effective in detecting structural alerts for all data sets, but mutagenicity, where eight descriptors detected mostly nitro groups. There are limited number of descriptors, which all three methods considered to be structural alerts. PNN tends to attribute descriptors to structural alerts that may be one of the reasons of its inferior efficiency compared to SVM. The described approach didn't allow detecting structural alerts for phospholipidosis. Though more than 30 descriptors were unanimously marked by the methods, all these descriptors refer to several groups of active compounds with similar structure (an example is demonstrated in Figure 3).

Performance of data visualization models

In this study, supervised extensions of Isomap and GTM were used for data visualization.

S-Isomap was first introduced in [53]. It demonstrated excellent results in separation different classes of training set. Mapping of the external test set is an important part of the chemography from the practical point of view in the context of the possibility of the application of the developed models to virtual screening and to mechanistic model interpretation which allows one to understand which changes of the existing structures are required to improve target properties, to generate new hypothesis and, finally, to optimize the chemical structures. In the original article for mapping an external test set it was recommended to use Radial Basis Network. In our study it turned out to be ineffective for diverse sets of chemical compounds. In this study, we propose new approach for the application of models to visualizing external data. We modified an approach proposed in [77,78] to adapt it for s-Isomap (See details in Method's description). The results of the mapping of external test sets for three types of activities are demonstrated in Figure 4. Hereinafter visualization maps are presented in the coordinate system generated be the applied methods. GTM and s-GTM presume that latent space is a rectangle of size 2×2 with its center located at (0, 0). Isomap and s-Isomap project compounds into two-dimensional space so that Euclidean distance (for Isomap) or dissimilarity measure (see Eq. 10) (for s-Isomap) can be preserved and the map scale is chosen accordingly. All axes in Figures 4, 5 and 6 are relative and have no units of measurement.

One can see that while s-Isomap performed almost perfect separation of the training set (none of the applied assessment parameters decreased below 0.91), the quality of mapping an external set for these models is highly dependent on the dataset in consideration. An external set of mutagenicity was mapped quite accurate

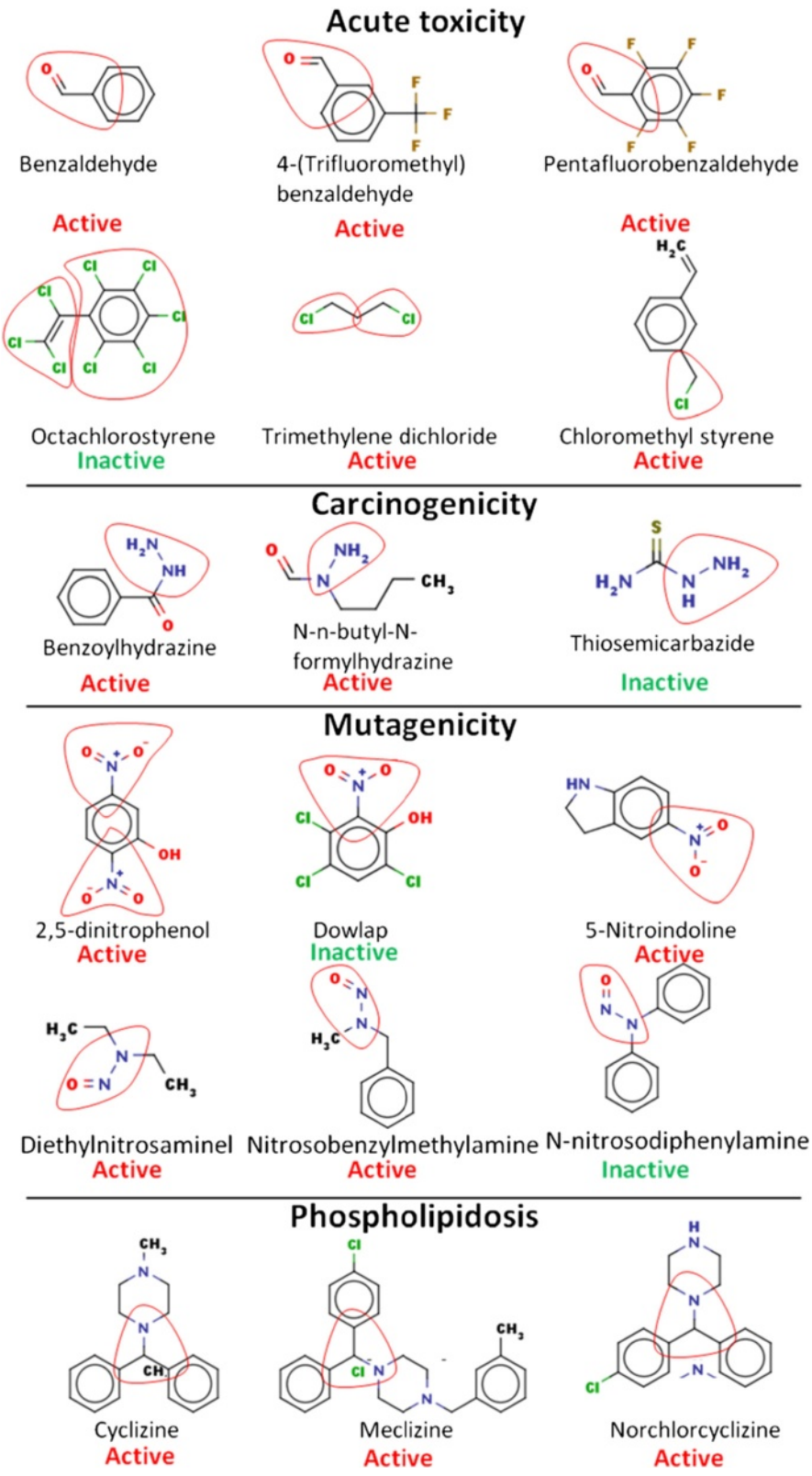

Figure 3 Examples of the descriptors frequently used to predict compounds as active by all three applied methods. For each descriptor an example of inactive compound is given (if any).

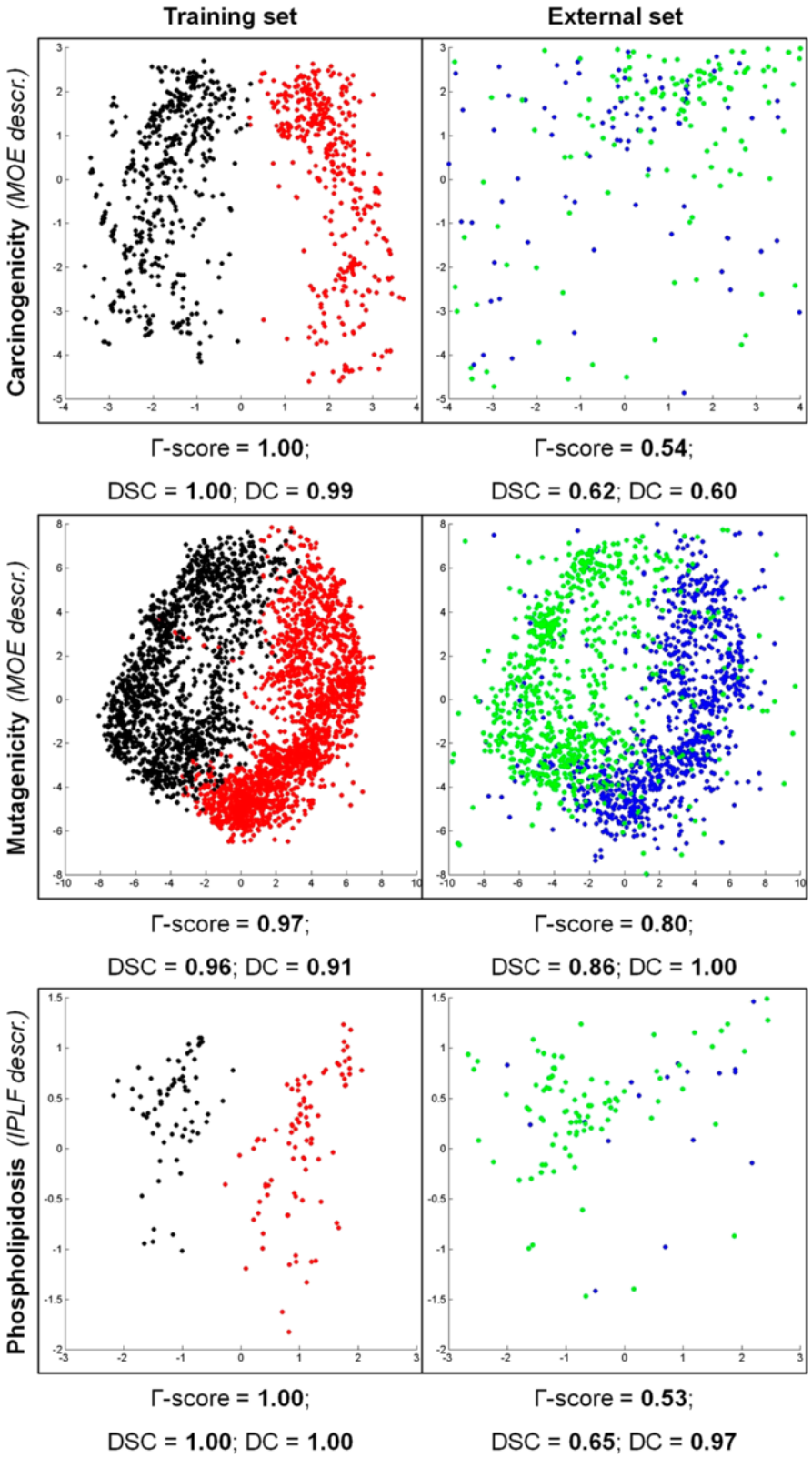

Figure 4 S-Isomap data visualization for considered datasets (the maps for the best combination of involved approach and descriptor type are given). Each point in the map corresponds to the individual compound (in red, blue - actives, black, green - inactives). In the left column the values of visualization quality assessment parameters are presented for the training set, in the right one – for the test set.

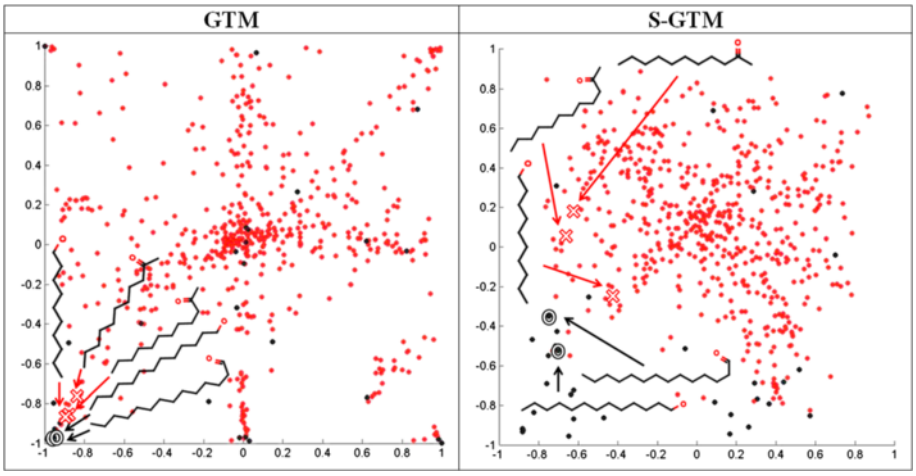

Figure 5 The visualization maps of acute toxicity dataset (MOE descriptors), obtained by unsupervised (left) and supervised (right) GTM. The singled out molecules share similar structure but belong to different classes. They are mapped close to each other by GTM, but are distinguishable in the s-GTM visualization.

(Γ-score = 0.80, DSC = 0.86, DC = 1.00), while the mapping of external set for carcinogenicity is moderate: the corresponding parameters varied in the range 0.54-0.62. One of the main factors that determine the quality of mapping is the distance from each point of the external set to the nearest neighbors in the training set. The closer they are, the better are the results. In case, if the distances are much greater than one, but are of the same scale, the additional parameter β_1 can be used to put them to the desirable range (look [53] for the specific values). In the case of carcinogenicity, in particular, the distances from the points of the external set differ for several degrees.

The supervised extension of GTM is proposed in this paper for the first time. It demonstrates a significant improvement in visualization performance. An example for acute toxicity dataset and MOE descriptors is given in Figure 5. Besides a noticeable increase in all three used visualization quality measures (Γ-score raised from 0.62 for unsupervised model to 0.77 for the supervised one, DSC – from 0.57 to 0.87 and DC – from 0.85 to 0.95, respectively), one can see how structurally similar compounds related to different classes and close to each other on the map obtained by unsupervised GTM are separated using supervised extension of GTM. Here, two groups were selected, each of them contained structurally similar active and inactive compounds. The first one contains toxic 1-Decanol and non-toxic 1-Tridecanol that differ from each other only by the length of the carbon chain (Tanimoto Similarity Coefficient (TSC) is equal to 1.00). The second group consists of toxic 2-Undecanone and 2-Dodecanone and similar to them (TSC = 0.82) non-toxic 3-Tetradecanal. All these compounds were mapped into a small area by unsupervised GTM while well distinguished applying its supervised extension.

Mapping of external test set for s-GTM is performed using the same procedure as for GTM, and the corresponding results are demonstrated in Figure 6. One can see that presented visualization maps are inferior to those of s-Isomap. At the same time s-GTM performs more accurate mapping of the external test set than s-Isomap, since after the model has been trained, the training set is mapped using the same algorithm as is used for the mapping of an external test set. In s-GTM, if one includes a compound from the training set in the test set, it will projected exactly to the same point of the map. This is not so for s-Isomap. Without label information each mapping will be an approximation and can be performed in different ways. The one we've proposed is based on the assumption that label information does not have much influence on the relative location of the points that are close to each other. During the training process s-Isomap changes distances between compounds in different manners regarding if the compounds belong to the same class or not but proportionally their relative position. Thus, new distances for compounds from different classes do not change significantly if they are close to each other. And if the compound from the test set has close neighbors in the training set, they will mapped close even if they belong to different classes. In Figure 6, as well as in Figure 4, acute toxicity maps are not presented since we had no corresponding external set at our disposal. Nevertheless, s-GTM demonstrated reasonably high results visualizing this data set. Considered quantitative measures for the best maps varied in the following ranges as a function of the descriptors type: Γ-Score – 0.76-0.77; DSC – 0.72-0.87; DC – 0.93-0.96.

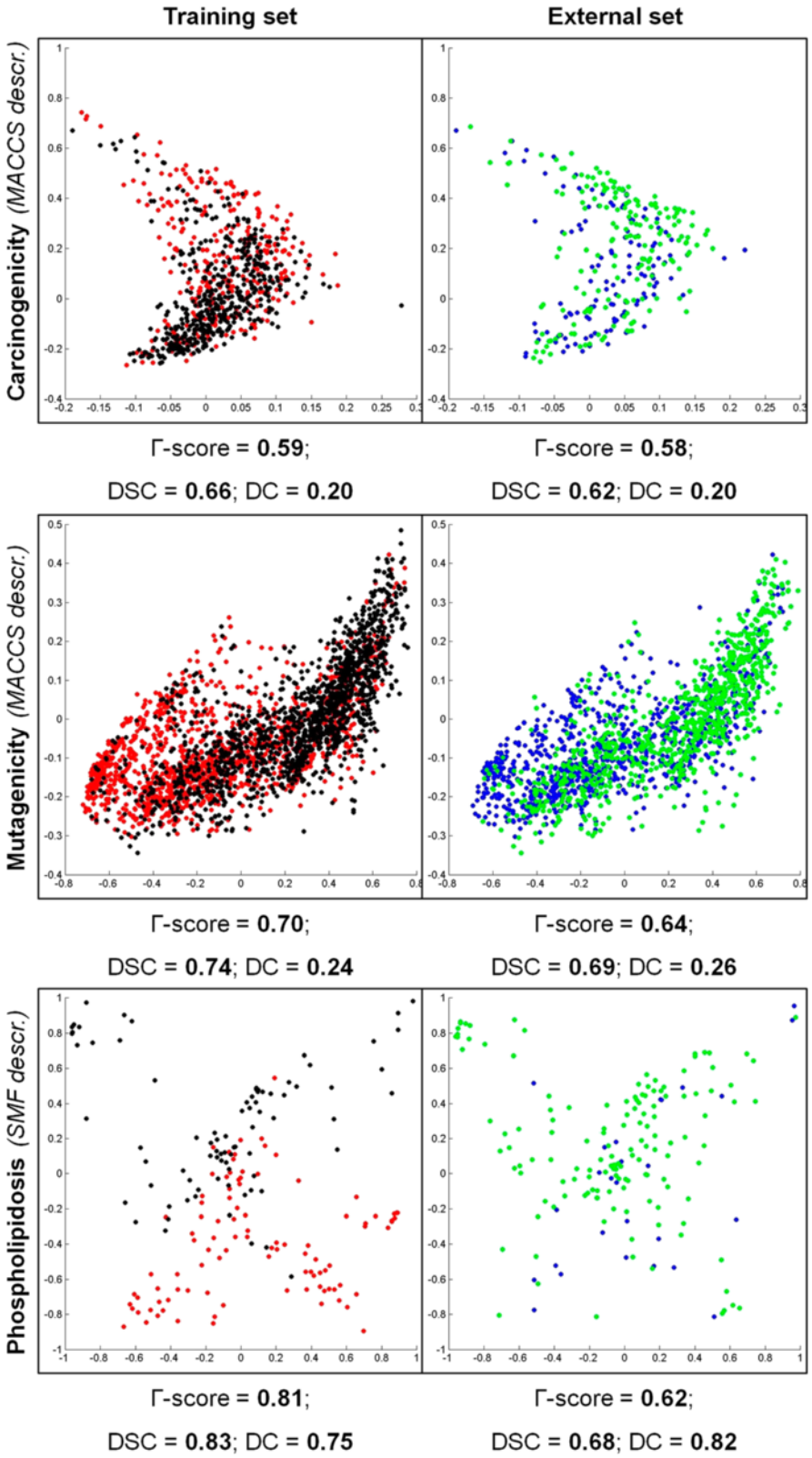

Figure 6 S-GTM data visualization for considered datasets (the maps for the best combination of involved approach and descriptor type are given). Each point in the map corresponds to the individual compound (in red, blue - actives, black, green - inactives). In the left column the values of visualization quality assessment parameters are presented for the training set, in the right one – for the test set.

The given examples allows one to assume that s-GTM tends to form clusters of identically labeled projections that is reflected by the increase of the DSC value as compared with the results of original GTM. For instance, for presented in Figure 6 examples the improvement in DSC is 0.12 for carcinogenicity, 0.18 for mutagenicity and 0.21 for phospholipidosis. At the same time, while generally s-GTM provides at least slight increase in all the considered parameters for visualization quality assessment, it doesn't separates areas of overlapping as successfully as s-Isomap does. The reason for this is that s-GTM works with the given relative location of compounds in the data space, while s-Isomap changes the distance between the compounds according to the label information (and thus performs some sort of metric learning [89]). E.g. if the choice of descriptors leads to overlapping differently labeled compounds in the original data space, s-GTM may not be able to separate them completely, but will project an external set following the pattern of the training set, while s-Isomap can achieve almost perfect separation for the most difficult visualization tasks, but then one may face some problems with the mapping of the external set.

For each presented map (Figures 4, 5 and 6) the values of three quantitative measures of visualization performances are given. None of the parameters is perfect and can be individually applied for identification of adequate data visualization models and comparison of different maps. Γ-score, for example, is high for the maps with randomly mixed compounds that are still grouped in small clusters. Distance consistency can be low for well separated classes that form non-convex figures. Distribution Consistency is usually high for imbalanced dataset visualization and strongly depends on its external parameter. The effectiveness of each parameter is defined by the nature of obtained map. For example, the maps may have similar DC value, but differ in DCS, which can be interpreted that considered maps have similar class overlapping and different level of clusterization. In this study, the combination of DC and DSC parameters

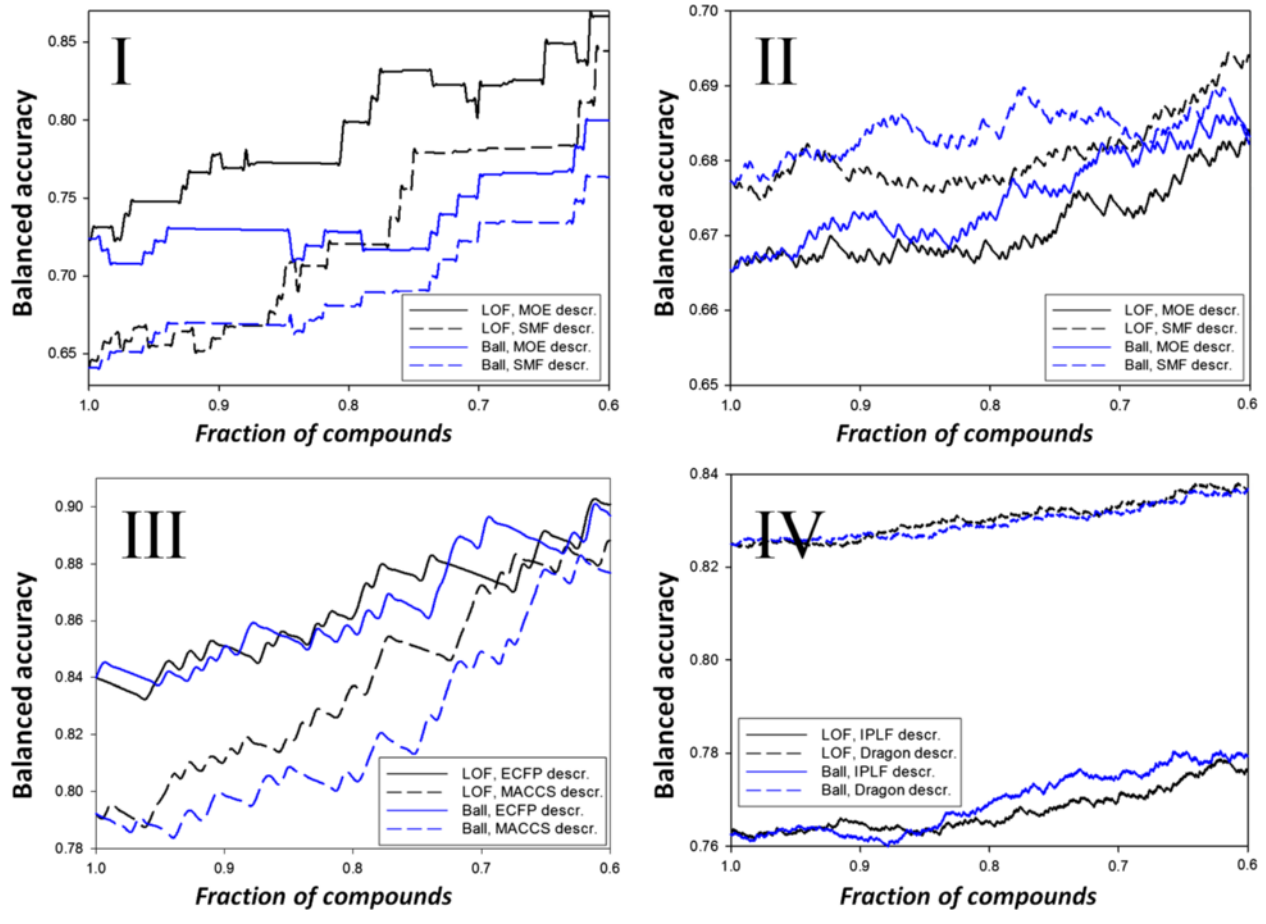

Figure 7 Balanced accuracy for SVM models as a function of data fraction. The compounds were removed according to the rate of their dissimilarity from the rest of the training set, which was assessed by the applicability domain methods. I – acute toxicity, II – carcinogenicity, III – phospholipidosis, IV - mutagenicity.

demonstrates its performance. Another advantage of DC and DSC is its less time- and memory-consuming compared to Γ-score.

Applicability domain of models

Two methods of applicability domain estimation were applied in this study, their performance was compared. One of them is a distance-based Ball, the other – a distribution-based LOF. The Principal Component Analysis was used as a pre-processing step. Each method was used to generate a sorted list of compounds according to their "outlierness" (the value of LOF function for LOF and distance to the centroid for ball). The impact of outliers' exclusion on the Balanced Accuracy of the models was analyzed.

In Figure 7 the Balanced Accuracy is given as a function of data fraction after the exclusion of outliers. The nature of the changes is affected by the distribution of compounds between classes in the dataset and predicting performance. In all the presented cases one can see a certain growth in performance which is different for considered datasets. Thus, the Balanced Accuracy of models for predicting carcinogenicity has increased only by 2.5% and after almost half of the compounds has been removed, while application of LOF to the SVM model for phospholipidosis that used SMF descriptors yield almost linear growth of Balanced Accuracy from 79% to 88% (after excluding 0.4 of compounds).

For acute toxicity LOF proved to be more efficient than *Ball*. This can be explained by the presence of several clusters with high density of compounds in the dataset containing compounds of different classes. The compounds in these clusters may have been correctly classified, while a number of false predictions were made for the compounds lying in the areas of classes overlapping in the midst of the clusters. In this case LOF was able to detect these mispredicted compounds as outliers and *Ball* just excluded the most distant from the centroid compounds in spite of the density distribution.

For phospholipidosis *Ball* and LOF demonstrated similar performance, though LOF is a bit more efficient. It may indicate that the data are slightly clusterized with an area of clusters' overlapping and most incorrectly predicted compounds are located far from the main aggregation of the chemical structures.

For carcinogenicity both applied methods demonstrated only a small increase of the Balanced Accuracy, with a better performance of *Ball* (in Figure 7 blue lines lie above corresponding black lines). This could happen if the projection of the dataset into the data space was a one cluster with irregular density distribution and large area of classes overlapping.

Similar pattern can be found for mutagenicity. Here, the maximum increase in BA is only about 2% and *Ball* only slightly outperforms LOF for IPLF descriptors. In respect with reasonable performances of both visualization

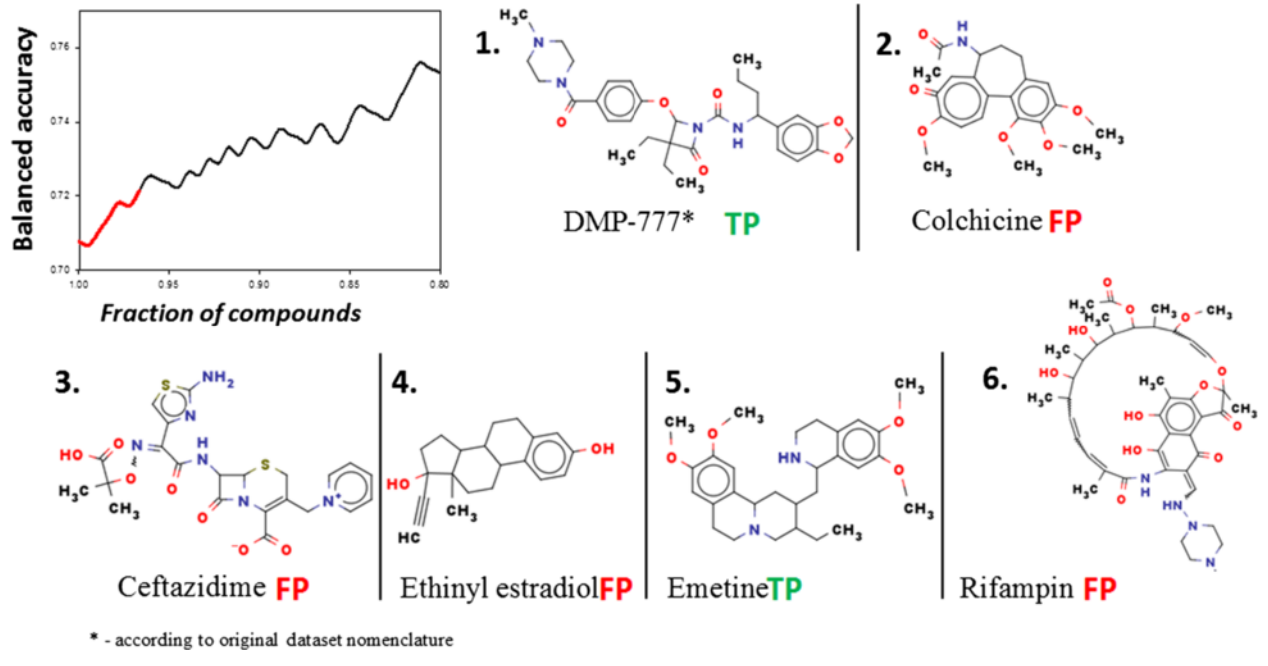

Figure 8 The examples of outliers, identified by LOF for combination of phospholipidosis and SMF descriptors and the influence of their exclusion on the Balanced Accuracy of a corresponding GTM model. The compounds given as examples are the ones that were marked as the most distant from the applicability domain. The red area on the diagram reflects the influence of their exclusion on the Balanced Accuracy of the model.

and classification methods for mutagenicity dataset, one may assume that this dataset doesn't contain many outliers and applying applicability domain analysis does not affect the predictive performance of models.

To demonstrate the principles of the outlier detection, the example of the compounds marked by LOF as the most dissimilar to the rest of the phospholipidosis data set was provided in Figure 8. The diagram on the upper left corner illustrates the effect of their exclusion on the Balanced Accuracy of the model obtained by GTM.

The SMF descriptors we used represent only terminal groups (See the section devoted to the descriptor types). The presented compounds were considered as outliers not because of the presence of some unique fragment, but because of unique or rare combination of atoms and bonds and their relative location. For example, Ceftazidime is the only compound in the dataset that contains sulfur with aromatic bond together with distanced heteroatoms (from 9 to 15 atoms in a fragment). And only in Rifampin there are carbon atoms with double bonds having from 4 to 10 atoms between them. Not all the given compounds are characterized by a number of unique descriptors, but all of them contain plenty of rare ones, as, for example, Colchicine.

Conclusions

This work concerns an approach that combines several classification and chemography methods for *in silico* assessment of chemical liabilities and for the interpretation of obtained results in the context of impact of structural changes of compounds on their pharmacological profile. Support Vector Machines, Generative Topographic Mapping and Probabilistic Neural Network were used for classification. The classification performances were improved by combination with two applicability domain assessment approaches (Ball and Local Outlier Factor), and their contribution was analyzed. Here, the supervised extension of Generative Topographic Mapping was proposed as new efficient chemography method. New approach for mapping new data using supervised Isomap without rebuilding models from the scratch has been proposed. The evaluation of the performance of the dimensionality reduction techniques and introduced descriptor spaces to separate different activity classes has been monitored by three parameters (Γ-score, Distance Consistency and Distribution Consistency) and their efficiency was compared. The obtained results, which are comparable with or exceed those, published by other teams for the given biological activities, allow one to use proposed approach as an efficient filter for exclusion of compounds with undesirable activities on early stages of drug design process.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AT provided regular supervisory input throughout the course of this work. NK conceived of the study, participated in its design, carried out calculations, and drafted the manuscript. SO participated in its design, carried out calculations, and drafted the manuscript. AB and ED carried out calculations. All authors read and approved the final manuscript.

Acknowledgments

Authors thank Russian Foundation for Basic Research (projects no. 11-03-00161 and 12-03-33086). Authors gratefully acknowledge Prof. Alexandre Varnek: the work on modification of Generative Topographic Mapping code was made in the Laboratory of Chemoinformatics in the University of Strasbourg under his supervision.

Author details

[1]Frumkin Institute of Physical Chemistry and Electrochemistry RAS, Leninsky pr-t 31-4, 119071 Moscow, Russia. [2]Moscow Institute of Physics and Technology, Institutsky per., 9, 141700 Dolgoprudny, Russia. [3]Kurnakov Institute of General and Inorganic Chemistry RAS, Leninsky pr-t 31, 119071 Moscow, Russia.

Received: 25 November 2013 Accepted: 28 April 2014
Published: 7 May 2014

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: **How to improve R&D productivity: the pharmaceutical industry's grand challenge.** *Nat Rev Drug Discov* 2010, **9:**203–214.
2. van de Waterbeemd H, Gifford E: **ADMET in silico modelling: towards prediction paradise?** *Nat Rev Drug Discov* 2003, **2:**192–204.
3. Oprea TI, Gottfries J: **Chemography: the art of navigating in chemical space.** *J Comb Chem* 2001, **3:**157–166.
4. Lee JA, Verleysen M: *Nonlinear Dimensionality Reduction.* New York: Springer; 2007.
5. Gorban AN, Kegl B, Wunsch DC, Zinovyev A: *Principal Manifolds for Data Visualisation and Dimension Reduction.* Berlin – Heidelberg – New York: Springer; 2007.
6. Ivanenkov YA, Bovina EV, Balakin KV: **Nonlinear mapping techniques for prediction of pharmacological properties of chemical compounds.** *Russ Chem Rev* 2009, **78:**465–483.
7. Ivanenkov YA, Savchuk NP, Ekins S, Balakin KV: **Computational mapping tools for drug discovery.** *Drug Discov Today* 2009, **14:**767–775.
8. Balakin KV: *Pharmaceutical Data Mining.* Wiley, New Jersey: Approaches and Applications for Drug Discovery; 2010.
9. Reutlinger M, Schneider G: **Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery.** *J Mol Graph Model* 2012, **34:**108–117.
10. Ertl P, Rohde B: **The Molecule Cloud-compact visualization of large collections of molecules.** *J Cheminform* 2012, **4:**1–8.
11. Ritchie TJ, Ertl P, Lewis R: **The graphical representation of ADME-related molecule properties for medicinal chemists.** *Drug Discov Today* 2011, **16:**65–72.
12. Jolliffe IT: *Principal Component Analysis.* New York: Springer; 2002.
13. Kruskal JB: **Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.** *Psychometrika* 1964, **29:**1–27.
14. Kruskal JB: **Nonmetric multidimensional scaling: a numerical method.** *Psychometrika* 1964, **29:**115–129.
15. Kohonen T: *Self-Organizing Maps.* Berlin: Springer-Verlag; 2001.
16. Agrafiotis DK, Xu H: **A self-organizing principle for learning nonlinear manifolds.** *Proc Natl Acad Sci U S A* 2002, **99:**15869–15872.
17. Agrafiotis DK: **Stochastic proximity embedding.** *J Comb Chem* 2003, **24:**1215–1221.
18. Rassokhin DN, Agrafiotis DK: **A modified update rule for stochastic proximity embedding.** *J Mol Graph Model* 2003, **22:**133–140.
19. Hinton GE, Roweis ST: **Stochastic Neighbor Embedding**. In *Advances in Neural Information Processing Systems.* Edited by Becker S, Thrun S, Obermayer K. Cambridge: The MIT Press; 2002:833–840.
20. Reutlinger M, Guba W, Martin RE, Alanine AI, Hoffmann T, Klenner A, Hiss JA, Schneider P, Schneider G: **Neighborhood-preserving visualization of adaptive structure–activity landscapes: application to drug discovery.** *Angew Chem Int Ed* 2011, **50:**11633–11636.

21. Sammon JW: **A nonlinear mapping for data structure analysis.** *IEEE T Comput* 1969, **18:**401–409.
22. Bishop CM, Svensen M: **GTM: the generative topographic mapping.** *Neural Comput* 1998, **10:**215–234.
23. Bishop CM, Svensén M, Williams CK: **GTM: A principled alternative to the self-organizing map**. In *Artificial Neural Networks — ICANN 96.* Edited by vor der Malsburg C, von Seelen W, Vorbrüggen JC, Sendhoff B. Berlin: Springer-Verlag; 1996:165–170.
24. Bishop CM, Svensén M, Williams CKI: **Developments of the generative topographic mapping.** *Neurocomputing* 1998, **21:**203–224.
25. Maniyar DM, Nabney IT, Williams BS, Sewing A: **Data visualization during the early stages of drug discovery.** *J Chem Inf Model* 2006, **46:**1806–1818.
26. Owen JR, Nabney I, Medina-Franco JL, Lopez-Vallejo F: **Visualization of molecular FIngerprints.** *J Chem Inf Model* 2011, **51:**1552–1563.
27. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A: **Generative Topographic Maps (GTM): universal tool for data visualization, structure-activity modeling and database comparison.** *Mol Inf* 2012, **31:**301–312.
28. Kireeva N, Kuznetsov SL, Bykov AA, Tsivadze AY: **Towards in silico identification of the human ether-a-go-go-related gene channel blockers: discriminative vs. generative classification models.** *SAR QSAR Environ Res* 2013, **24:**103–117.
29. Kireeva N, Kuznetsov SL, Tsivadze AY: **Toward navigating chemical space of ionic liquids: prediction of melting points using generative topographic maps.** *Ind Eng Chem Res* 2012, **51:**14337–14343.
30. Hasegawa K, Funatsu K: **Prediction of protein-protein interaction pocket using L-Shaped PLS approach and its visualizations by generative topographic mapping.** *Mol Inf* 2014, **33:**65–72.
31. Hähnke V, Rupp M, Krier M, Rippmann F, Schneider G: **Pharmacophore alignment search tool: influence of canonical atom labeling on similarity searching.** *J Comb Chem* 2010, **31:**2810–2826.
32. Das P, Moll M, Stamati H, Kavraki LE, Clementi C: **Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction.** *Proc Natl Acad Sci* 2006, **103:**9885–9890.
33. Chen N, Lu W, Yang J, Li G: *Support vector machine in chemistry.* Singapore: World Scientific; 2004.
34. Specht DF: **Probabilistic neural networks.** *Neural Netw* 1990, **3:**109–118.
35. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T: **QSAR applicability domain estimation by projection of the training set descriptor space: a review.** *ALTA Altern Lab Anim* 2005, **33:**445–459.
36. Tetko IV, Bruneau P, Mewes H-W, Rohrer DC, Poda GI: **Can we estimate the accuracy of ADMET predictions?** *Drug Discov Today* 2006, **11:**700–707.
37. Weaver S, Gleeson MP: **The importance of the domain of applicability in QSAR modeling.** *J Mol Graph Model* 2008, **26:**1315–1326.
38. Todeschini R, Consonni V, Pavan M: **A distance measure between models: a tool for similarity/diversity analysis of model populations.** *Chemometr Intell Lab* 2004, **70:**55–61.
39. Schultz TW, Hewitt M, Netzeva TI, Cronin MT: **Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action.** *QSAR Comb Sci* 2007, **26:**238–254.
40. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller K-R: **Applicability domains for classification problems: benchmarking of distance to models for AMES mutagenicity set.** *J Chem Inf Model* 2010, **50:**2094–2111.
41. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A: **Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection.** *J Chem Inf Model* 2008, **48:**1733–1746.
42. Soto AJ, Vazquez GE, Strickert M, Ponzoni I: **Target-driven subspace mapping methods and their applicability domain estimation.** *Mol Inf* 2011, **30:**779–789.
43. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R: **Comparison of different approaches to define the applicability domain of QSAR models.** *Molecules* 2012, **17:**4791–4810.
44. Rodgers A, Zhu H, Fourches D, Rusyn I, Tropsha A: **Modeling liver-related adverse effects of drugs using k nearest neighbor quantitative structure-activity relationship method.** *Chem Res Toxicol* 2010, **23:**724–732.
45. Sheridan RP: **Three useful dimensions for domain applicability in QSAR models using random forest.** *J Chem Inf Model* 2012, **52:**814–823.
46. Sahigara F, Ballabio D, Todeschini R, Consonni V: **Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions.** *J Cheminform* 2013, **5:**27.
47. Todeschini R, Ballabio D, Consonni V, Sahigara F, Filzmoser P: **Locally centred Mahalanobis distance: a new distance measure with salient features towards outlier detection.** *Anal Chim Acta* 2013, **787:**1–9.
48. Tetko IV, Novotarskyi S, Sushko I, Ivanov V, Petrenko AE, Dieden R, Lebon F, Mathieu B: **Development of dimethyl sulfoxide solubility models using 163 000 molecules: using a domain applicability metric to select more reliable predictions.** *J Chem Inf Model* 2013, **53:**1990–2000.
49. Brandmaier S, Novotarskyi S, Sushko I, Tetko IV: **From descriptors to predicted properties: experimental design by using applicability domain estimation.** *ATLA Altern Lab Anim* 2013, **41:**33–47.
50. Tax D: *Data description toolbox dd tools 1.7. 5.* Delft: Delft University of Technology; 2010.
51. Breunig MM, Kriegel H-P, Ng RT, Sander J: **LOF: identifying density-based local outliers.** *ACM Sigmod Record* 2000, **29:**93–104.
52. Kireeva N, Ovchinnikova S, Tsivadze A: **Supervised Generative Topographic Mapping for In Silico Assessment of Chemical Liabilities**. In *Proceedings of ACS National Meeting "Chemistry in Motion" Indianapolis.* 2013.
53. Geng X, Zhan D-C, Zhou Z-H: **Supervised nonlinear dimensionality reduction for visualization and classification.** *IEEE T Syst Man Cy B* 2005, **35:**1098–1107.
54. Tropsha A: **Best practices for QSAR model development, validation, and exploitation.** *Mol Inf* 2010, **29:**476–488.
55. **Chemaxon Standardizer.** http://www.chemaxon.com/products/standardizer/.
56. **Instant JChem.** http://www.chemaxon.com/products/instant-jchem/.
57. Kazius J, McGuire R, Bursi R: **Derivation and validation of toxicophores for mutagenicity prediction.** *J Med Chem* 2005, **48:**312–320.
58. **DSSTox database.** http://www.epa.gov/ncct/dsstox/.
59. Lowe R, Mussa HY, Nigsch F, Glen RC, Mitchell JB: **Predicting the mechanism of phospholipidosis.** *J Cheminform* 2012, **4:**2.
60. Goracci L, Ceccarelli M, Bonelli D, Cruciani G: **Modeling phospholipidosis induction: reliability and warnings.** *J Chem Inf Model* 2013, **53:**1436–1446.
61. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA: **Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (Pimephales promelas).** *Environ Toxicol Chem* 1997, **16:**948–967.
62. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G: **ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors.** *Curr Comput Aided Drug Des* 2008, **4:**191–198.
63. Ruggiu F, Marcou G, Varnek A, Horvath D: **ISIDA property-labelled fragment descriptors.** *Mol Inf* 2010, **29:**855–868.
64. **Molecular Operating Environment.** www.chemcomp.com.
65. Guha R: **Chemical informatics functionality in R.** *J Stat Softw* 2007, **18:**1–16.
66. **R project.** http://www.r-project.org/foundation/.
67. **Dragon 6.** http://www.talete.mi.it/products/dragon_molecular_descriptors.htm.
68. Cristianini N, Shawe-Taylor J: *An Introduction To Support Vector Machines (and Other Kernel-Based Learning Methods).* Cambridge: Cambridge University Press; 2000.
69. Ivanciuc O: *Applications of Support Vector Machines in Chemistry.* Weinheim: Wiley-VCH; 2007.
70. Vapnik VN: *Statistical Learning Theory.* New York: Wiley-Interscience; 1998.
71. Vapnik VN: *The Nature of Statistical Learning Theory.* New York: Springer-Verlag; 1995.
72. Bishop CM: *Pattern Recognition and Machine Learning.* New York: Springer; 2006.
73. Pearl J: **Bayesian networks: a model of self-activated memory for evidential reasoning.** In *Proceedings of The 7th conference of the Cognitive Science Society.* University of California, Irvine; 1985:329–334.
74. Hand D: *Kernel discriminant analysis*, Research studies press Chichester. 1982.
75. Tenenbaum JB, De Silva V, Langford JC: **A global geometric framework for nonlinear dimensionality reduction.** *Science* 2000, **290:**2319–2323.
76. Mardia KV, Kent JT, Bibby JM: *Multivariate Analysis.* London: Academic Press; 1979.
77. Bengio Y, Paiement J-F, Vincent P, Delalleau O, Le Roux N, Ouimet M: **Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering**. In *Advances in Neural Information Processing Systems.* Edited by Thrun S, Saul LK, Scholkopf B. Cambridge: MIT Press; 2004:177–184.
78. Silva VD, Tenenbaum JB: **Global versus local methods in nonlinear dimensionality reduction**. In *Advances in Neural Information Processing*

Systems. Edited by Becker S, Thrun S, Obermayer K. Cambridge: MIT Press; 2002:705–712.
79. Sokolova M, Japkowicz N, Szpakowicz S: **Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation**. In *Advances in Artificial Intelligence*. Edited by Sattar A, Kang BH. New York: Springer; 2006:1015–1021.
80. Chang CC, Lin CJ: **LIBSVM: a Library for Support Vector Machines.** http://www.csie.ntu.edu.tw/~cjlin/libsvm.
81. Nabney I, Bishop C: **Netlab neural network software.** http://ntlb.sourceforge.net/.
82. Stork DG, Yom-Tov E: *Computer Manual in MATLAB to Accompany Pattern Classification*. New York: John Wiley & Sons; 2004.
83. Sips M, Neubert B, Lewis JP, Hanrahan P: **Selecting good views of high-dimensional data using class consistency.** *Comput Graph Forum* 2009, **28:**831–838.
84. Singh KP, Gupta S, Rai P: **Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches.** *Ecotox Environ Safety* 2013, **95:**221–233.
85. Öberg T: **A QSAR for baseline toxicity: validation, domain of application, and prediction.** *Chem Res Toxicol* 2004, **17:**1630–1637.
86. Cassani S, Kovarich S, Papa E, Roy PP, van der Wal L, Gramatica P: **Daphnia and fish toxicity of (benzo) triazoles: validated QSAR models, and interspecies quantitative activity–activity modelling.** *J Hazard Mater* 2013, **258:**50–60.
87. Devillers J, Mombelli E, Samsera R: **Structural alerts for estimating the carcinogenicity of pesticides and biocides.** *SAR QSAR Environ Res* 2011, **22:**89–106.
88. Liu R, Wallqvist A: **Merging applicability domains for in silico assessment of chemical mutagenicity.** *J Chem Inf Model* 2014, **54:**793–800.
89. Kireeva NV, Ovchinnikova SI, Kuznetsov SL, Kazennov AM, Tsivadze AY: **Impact of distance-based metric learning on classification and visualization model performance and structure–activity landscapes.** *J Comput Aid Mol Des* 2014, **28:**61–73.

doi:10.1186/1758-2946-6-20
Cite this article as: Ovchinnikova *et al.*: Supervised extensions of chemography approaches: case studies of chemical liabilities assessment. *Journal of Cheminformatics* 2014 **6**:20.