J Comput Aided Mol Des (2014) 28:61–73
DOI 10.1007/s10822-014-9719-1

# Impact of distance-based metric learning on classification and visualization model performance and structure–activity landscapes

**Natalia V. Kireeva · Svetlana I. Ovchinnikova · Sergey L. Kuznetsov · Andrey M. Kazennov · Aslan Yu. Tsivadze**

Received: 15 March 2013 / Accepted: 24 January 2014 / Published online: 4 February 2014


**Abstract** This study concerns large margin nearest neighbors classifier and its multi-metric extension as the efficient approaches for metric learning which aimed to learn an appropriate distance/similarity function for considered case studies. In recent years, many studies in data mining and pattern recognition have demonstrated that a learned metric can significantly improve the performance in classification, clustering and retrieval tasks. The paper describes application of the metric learning approach to in silico assessment of chemical liabilities. Chemical liabilities, such as adverse effects and toxicity, play a significant role in 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. Here, to our knowledge for the first time, a distance-based metric learning procedures have been applied for in silico assessment of chemical liabilities, the impact of metric learning on structure–activity landscapes and predictive performance of developed models has been analyzed, the learned metric was used in support vector machines. The metric learning results have been illustrated using linear and non-linear data visualization techniques in order to indicate how the change of metrics affected nearest neighbors relations and descriptor space.

N. V. Kireeva · S. I. Ovchinnikova · S. L. Kuznetsov · A. Yu. Tsivadze
Frumkin Institute of Physical Chemistry and Electrochemistry RAS, Leninsky Prospect, 31a, 119071 Moscow, Russia

N. V. Kireeva (✉) · S. I. Ovchinnikova · A. M. Kazennov
Moscow Institute of Physics and Technology, Institutsky Per., 9, 141700 Dolgoprudny, Russia
e-mail: nkireeva@gmail.com; kireeva@phyche.ac.ru; nkireeva@unistra.fr

**Keywords** Chemical liabilities · Large margin nearest neighbors · Metric learning · Structure–activity landscapes · Chemography · Generative topographic maps

## Introduction

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. It allows one 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. 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.

In recent years, many studies in data mining and pattern recognition demonstrated that learned metric (ML) can significantly improve the performance in classification, clustering and retrieval tasks [3–7]. In chemoinformatics, many popular machine learning techniques, for instance, K-means, nearest-neighbors classifiers and SVM, significantly rely on the distance/similarity function to reproduce correctly the relationships in chemical data. This study concerns Large Margin Nearest Neighbors approach [8] and its multi-metric extension as the efficient methods for metric learning which aimed to learn an appropriate distance/similarity function for the defined specific objective and their application for in silico assessment of chemical liabilities of chemical compounds.

*K*-nearest neighbors (*k*NN) approach [9] is one of the oldest and simplest classification methods. Despite its

simplicity, the *k*NN rule often yields competitive results or can improve the state-of-the-art results. The performance of *k*NN classification depends crucially on the distance/similarity metric used to identify the nearest neighbors. Most of *k*NN classifiers use by default the Euclidean distance to measure the similarity between compounds represented by their descriptor vectors. In the same time, the distance metric for *k*NN classification can be adapted to the specific problem being solved. The endpoint of the metric learning procedure for *k*NN classification is to group the compounds of the training set with the same property label together while the compounds from different classes should be separated by a large margin. Recently, a number of researchers in machine learning and pattern recognition have demonstrated that *k*NN classification can be significantly improved by learning an appropriate distance metric from labeled examples [3, 4, 7].

The concept of molecular similarity is widely used in medicinal chemistry and chemoinformatics. The principal idea is that structurally more similar molecules are more likely to exhibit similar properties than structurally less similar molecules. Accordingly, choosing appropriate metrics for the estimation of similarity between molecules is a key point for this approach to work [10–15]. One of the most known techniques of metric learning published in chemoinformatics is realized in ASNN software, where *k*NN is used to improve the results of the neural networks introducing the correction in the space of models [16, 17].

Here, to our knowledge for the first time, the impact of a distance-based metric learning on Structure–activity landscapes has been analyzed. Structure–activity landscapes [18–20] are constructed from chemical spaces by adding an additional dimension associated with activity value. Several types of features can be considered in activity landscapes: activity cliffs, similarity cliffs, smooth SAR regions and featureless regions [21]. Activity cliffs considered in this study are the pairs of structurally similar compounds with significantly different activities. Analysis of activity cliffs can improve the understanding in which regions of chemical space the model is more or less predictive. Assessment of structure–activity cliffs is carried out by pairwise comparisons of compounds and their related activity values. Guha and Van Drie recently proposed the use of Structure Activity Landscape Index (SALi) [22, 23] as a measure to assess structure–activity landscapes, where SALi is introduced as the ratio of the absolute value of the activity difference to the distance in the chemical space between pair of compounds. Thus, a small SALi value is indicative of a smooth activity transition, whereas a large one corresponds to the presence of an activity cliff.

There are two main goals in the current work. First, we consider how the metric learning impacts on predictive performance of developed models. The second goal of this study is an assessment of impact of metric learning on structure–activity landscapes. Here, to our knowledge for the first time, a distance-based metric learning procedure has been applied for in silico assessment of chemical liabilities, the impact of metric learning on classification and visualization models' performance and structure–activity landscapes has been analyzed. The impact of principal component analysis as a pre-processing step of model development has been analyzed. The learned metric has been used in support vector machines. The results of proposed approaches have been compared with those obtained using ASNN software. Metric learning has been carried out for four data sets: (1) a set of 242 compounds with measured $pIC_{50}$ values for hERG inhibition [24], (2) carcinogenicity data collected from the public toxicity database network (DSSTox) and containing 1,088 chemical structures [25], (3) a set of 181 compounds measured on inducing phospholipidosis [26] and (4) a set of 882 direct factor Xa inhibitors [27].

## Methods

### Data and descriptors

Four data sets were considered in this study. A set of 242 $pIC_{50}$ values for hERG inhibition was taken from [24]. To generate the classification models the considered data set was split into two classes according to their activities on the hERG channel inhibition. The $pIC_{50} = 5$ (low micromolar potency) was considered as the threshold value for hERG inhibition. Thus, 104 inactive and 138 active compounds for hERG channel inhibition have been involved in model development.

A set of 100 phospholipidosis-inducing compounds and 82 negative druglike compounds were taken from [26], 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).

Carcinogenicity data was collected from the distributed ISSCAN Database (part of structure-searchable toxicity DSSTox public database network [25]). 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.

An additional structure–activity dataset of direct factor Xa inhibitors was taken from [27]. It contains information on 882 compounds.

The data preparation has been carried out using recommendations published in [28]. Chemaxon Standardizer [29] and Instant JChem [30] software have been used for the data preparation. Using Standardizer, the explicit hydrogen atoms have been removed, the structures have been aromatized. ISIDA substructural molecular fragments (SMF) [31] have been used as descriptors in this study. SMF are subgraphs of a molecular graph, whereas their occurrences are the descriptor values. The subclass of the SMF descriptors consisting of the shortest topological paths with explicit representation of only terminal atoms and bonds was used, where the values of minimal $n_{min}$ and maximal $n_{max}$ number of atoms varied from 2 to 15. The Floyd algorithm [32] was applied for finding the shortest paths in the molecular graphs. Single, double, triple and aromatic bonds were recognized.

## Metric learning

### *Large margin nearest neighbors approach*

The large margin nearest neighbors (LMNN) algorithm was proposed in [8]. It has been developed to improve performance of $k$-nearest neighbors ($k$NN) classifier. LMNN is based on the assumption, that $k$NN predicts class membership of each compound more accurate if all its $k$ nearest neighbors are related to the same class. The more training points will be surrounded by neighbors with the same class label, the higher possibility that $k$ nearest neighbors of each test point will also share the same class label. Thus, the algorithm attempts to find a linear transformation $x' = \mathbf{L}x$ of the input space that moves neighbors with corresponding class labels closer and increase distances between compounds from different classes.

At first, target neighbors for each compound are identified (computing the $k$ nearest neighbors with the same class label using Euclidean distance). They are fixed a priori and do not change during the learning process. Each pair of a training point $x_i$ and its target neighbor $x_j$ defines a neighborhood area, which is a sphere with the center in $x_i$ and radius $r = |x_i - x_j|$. Any $x_l$ that falls in the neighborhood area of differently labeled compound plus unit margin is called an imposter.

Thereby the learning process pursues two goals. The first one is to shrink neighborhood areas and the second is to move impostors out of them, creating a finite margin between neighborhoods and impostors (Fig. 1). They are represented in the loss function $\varepsilon(\mathbf{L})$, which consists of two terms. The first one

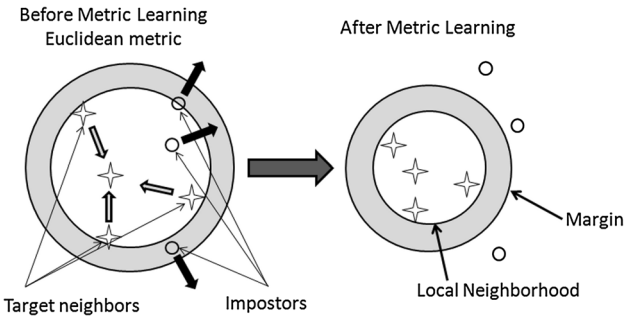


**Fig. 1** Schematic illustration of Large Margin Nearest Neighbors algorithm. The algorithm attempts to group the compounds of the training set with the same property label together while the compounds from different classes should be separated by a large margin Here, the distance metric is optimized so that: (1) its $k = 3$ target neighbors lie within a smaller radius after training; (2) differently labeled inputs lie outside this smaller radius by some margin. This presentation is inspired by figure in the original publication by Weinberger et al. [8]

$$\varepsilon_{pull}(\mathbf{L}) = \sum_{ij} \eta_{ij} \left\| \mathbf{L}(x_i - x_j) \right\|^2 \quad (1)$$

penalizes big distances between compounds sharing the same class label which are recognized as $k$-nearest neighbors, while the second term

$$\varepsilon_{push}(\mathbf{L}) = \sum_{ijl} \eta_{ij}(1 - y_{il}) \left[ 1 + \left\| \mathbf{L}(x_i - x_j)^2 \right\| - \left\| \mathbf{L}(x_i - x_l)^2 \right\| \right]_+ \quad (2)$$

penalizes small distances between differently labeled compounds.

Here $\eta_{ij}$ is used to indicate whether $x_j$ is a target neighbor of $x_i$ ($\eta_{ij} \in \{0, 1\}$), $y_{ij}$ is used to indicate whether or not the label $y_i$ and $y_j$ match.

A weighting parameter $\mu \in [0, 1]$ balances these goals:

$$\varepsilon(\mathbf{L}) = (1 - \mu)\varepsilon_{pull}(\mathbf{L}) + \mu\varepsilon_{push}(\mathbf{L}) \quad (3)$$

Minimizing this function yields a linear transformation of the input space to the space adapted for the case study. The transformation defines a new pseudometric (pseudometric is metric that does not satisfy the distinguishability property, i.e. zero distance between two points $x_i$ and $x_j$ does not necessarily mean their identity) with squared distances

$$D(x_i, x_j) = \left\| \mathbf{L}(x_i - x_j) \right\|_2^2 \quad (4)$$

They can be also expressed in terms of square matrix $\mathbf{M} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$:

$$\mathcal{D}_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j). \quad (5)$$

Such pseudometric can be considered as a generalization of Mahalanobis metric (in which $\mathbf{M}$ is inversed covariance matrix for Gaussian distribution). If $\mathbf{L}$ is a real-value

matrix, then **M** is always positive semidefinite. So we can estimate either linear transformation **L** or positive semidefinite **M**. The second approach allows us to apply the semidefinite programming (SDP) [33].

In the terms of matrix **M** Eq. (3) can be rewritten as

$$\varepsilon(\mathbf{M}) = (1-\mu)\sum_{ij}\eta_{ij}\mathcal{D}_{\mathbf{M}}(x_i,x_j) + \mu\sum_{ijl}\eta_{ij}(1-y_{il})\left[1+\mathcal{D}_{\mathbf{M}}(x_i,x_j)-\mathcal{D}_{\mathbf{M}}(x_i,x_l)\right]_+ \quad (6)$$

where **M** is positive semidefinite.

This loss function is piecewise linear and convex of the elements in the matrix **M**.

To obtain SDP we need to introduce slack variables $\xi_{ijl}$ for each triple of target neighbors ($x_j$ is a target neighbor of $x_i$) and impostors $x_l$. These non-negative variables represent the amount of margin violation of an impostor with respect to any pair of a compound and its target neighbor. Using them one can obtain the following SDP

$$Minimize\ (1-\mu)\sum_{ij}\eta_{ij}(x_i-x_j)^{\mathrm{T}}\mathbf{M}(x_i-x_j) + \mu\sum_{ijl}\eta_{ij}(1-y_{ij})\xi_{ijl}\quad subject\ \ to:$$

$$(1)\ (x_i-x_l)^{\mathrm{T}}\mathbf{M}(x_i-x_l)-(x_i-x_j)^{\mathrm{T}}\mathbf{M}(x_i-x_j)\geq 1-\xi_{ijl}$$

$$(2)\ \xi_{ijl}\geq 0$$

$$(3)\ \mathbf{M}\succcurlyeq 0$$

LMNN constructs new global metric, i.e. metric in transformed space is completely defined by a single matrix **M** (or **L**). But it is also may be useful to learn metric locally so that in different regions of the input space can be applied different metrics. This opportunity is provided in the so called multi-metric modification of LMNN.

### *Multi-metric LMNN (MLMNN)*

There is an extension of LMNN [8] developed to replace the global linear transformation of the input space in the multiple local ones that could be especially effective for large diverse data sets. It can be performed dividing the training data into separate clusters using k-means, spectral (or using alternative clustering algorithms) or dividing into classes using the label information as it has been done in this study (one metric for each class). New metric is learned for each class individually. The latter is achieved by solving the modified version of SDP algorithm, where multiple metrics are learned simultaneously by solving a single SDP. It allows one to guarantee the same scale of the distances since the distances computed in different scales could not be compared and used for kNN classification.

## Data visualization techniques

In this study, we use two representatives of dimensionality reduction methods related to two different families: linear (Principal Component Analysis PCA [34]) and nonlinear (Generative Topographic Mapping GTM [35, 36]). PCA is probably the best-known technique for dimensionality reduction. GTM recently has been efficiently used in chemoinformatics in a number of studies [37–42].

## Computational procedures

The basic implementation of LMNN and MLMNN approaches have been taken from [43]. The models have been developed for two cases: with and without use of Principal Component Analysis (PCA) [34] as a pre-processing step. The latter was recommended in [8] in order to reduce the dimensionality of the inputs.

### *Grid optimization of internal metric learning parameters*

There are several parameters to be adjusted: (1) number of principal components (depending on data set) and (2) weighting parameter μ (0.2, 0.5 and 0.8) (see Sect. "Metric Learning"). Since the discriminatory power of the approaches in classification was considered as a quality criterion, the abovementioned parameters were tuned using internal test set. The predictive performance of developed classification models were assessed using three-fold external cross-validation (3-CV) procedure considering balanced accuracy (BA) value [44] as a criterion of the predictive performance of the models.

The performance of data visualization has been monitored using *Γ-score* parameter introduced in [38]:

$$\Gamma_{M,D}(k) = \frac{1}{n}\sum_{i=1}^{n}\gamma_i(k);$$

$$\text{where } \gamma(k) = \frac{1}{N_v}\sum_{1}^{N_v}G(i,k) \text{ and } G(i,k) = \frac{1}{k}\sum_{j=1}^{k}g(v_i,j);$$

Here, $k$—number of nearest neighbors; $N_v$—number of examples; $g(v_i, j) = 1$ if the molecule that is $j$-th nearest to $v_i$ in the visualization space is in the same class, $g(v_i, j) = 0$ otherwise.

This score characterizes the ability of a model to produce similar-structure clustering in a visualization and can be computed for a data set where the information about the classes is available. The greater the value of *Γ-score*, the better the separation of the classes in the visualization.

In order to check if the learned metric with parameters optimized for *k*NN will work with other machine learning

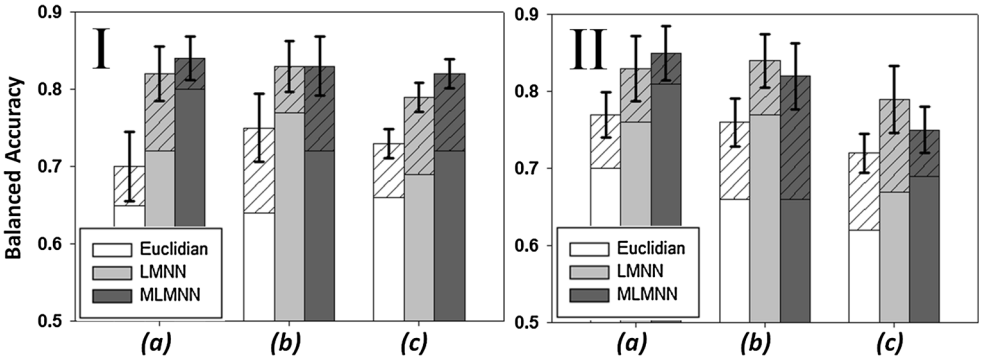


**Fig. 2** Distance Metric Learning with **(I)** or without **(II)** PCA for pre-processing: *Balanced Accuracy* of classification for *a* HERG, *b* phospholipidosis and *c* carcinogenicity data sets. The hatched areas demonstrate the increase of BA after application of bootstrap procedure

approaches as well we have used the linear transformation function L to transfer the optimized metric in SVM. SVM classification has been performed using LIBSVM package [45] with RBF kernel function. Two parameters of the method, ν and γ, have been varied within the following ranges: $\nu = 0.001, 0.051,\ldots, 0.91$ (internal parameter of method) and $\gamma = 2^{-15}, 2^{-13},\ldots, 2^{3}$ (parameter of RBF kernel) for fivefold cross-validation procedure.

For calculation of statistical significance of the obtained classification results the bootstrap (as its implementation a standard function of Matlab Software was taken) was used. 100 random samples were generated out of each dataset and for each of them the values of Balanced Accuracy were assessed, what allowed us to find the confidence intervals and to measure *p* values according to Student's t-distribution.

## Results and analysis

### Predictive performance of developed classification models

LMNN classification has been evaluated on four data sets of varying size and difficulty. Information on the parameters choice for the applied methods can be found in Sect. "Impact of parameters choice" In Fig. 2 (left) the BA of the models is represented as a function of metrics in use for *k*NN classification using PCA as pre-processing step for model development. Figure 2 (right) show obtained results without PCA application. One can see that in most of cases there is no difference in the results performance using PCA or not. Thus, the main reason for using PCA is related to the efficiency of solving the SDP since its application has no impact on the predictive performance of the models or on model interpretation. For the models developed with PCA BA value varies as a function of data set and increased from 0.65 in initial Euclidean metrics to 0.72 using LMNN for metric learning and to 0.8 using multiple metrics (MLMNN) for hERG; from 0.64 to 0.77 (LMNN) and to 0.72 (MLMNN) for phospholipidosis and from 0.66 to 0.69 (LMNN) and to 0.72 (MLMNN) for carcinogenicity data sets. Without PCA application the similar trend in performance improvement is observed: BA value increased from 0.7 in initial Euclidean metrics to 0.76 using LMNN for metric learning and to 0.81 using MLMNN for hERG; from 0.66 to 0.77 (LMNN) and to 0.65 (MLMNN) for phospholipidosis and from 0.62 to 0.67 (LMNN) and to 0.69 (MLMNN) for carcinogenicity data sets.

One can see, that in most of the cases metric learning leads to improvement of the results except of applying multiple metrics extension of LMNN for phospholipidosis data set whereas for the same data set LMNN application resulted in significant improvement of model performance. In [8] the assumption about the dependence of the results performance from the data density has been made in the discussion about the difference in the results quality for several data sets derived from collections of images, speech, and text. This assumption has been examined in our study by computing the Euclidean distance, Mahalanobis distance and Jaccard coefficient (averaged over all compound pairs) for considered data sets and has not been confirmed.

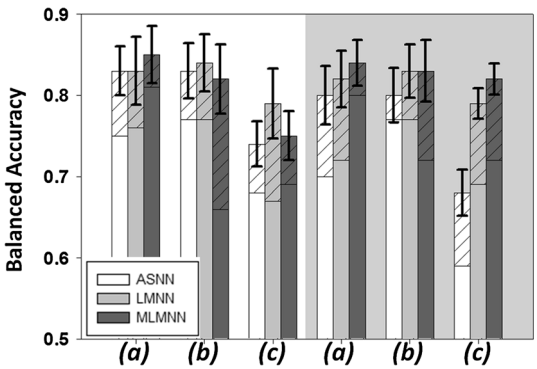


**Fig. 3** The comparison between ASNN, LMNN and MLMNN. *Balanced Accuracy* of classification for *a* HERG, *b* phospholipidosis and *c* carcinogenicity data sets. Three groups of *bars* on the *left* (on the *white background*) show the classification results without using PCA as a preprocessing step and three groups on the *right* (on the *grey background*)—with PCA. The hatched areas demonstrate the increase of BA after application of bootstrap procedure

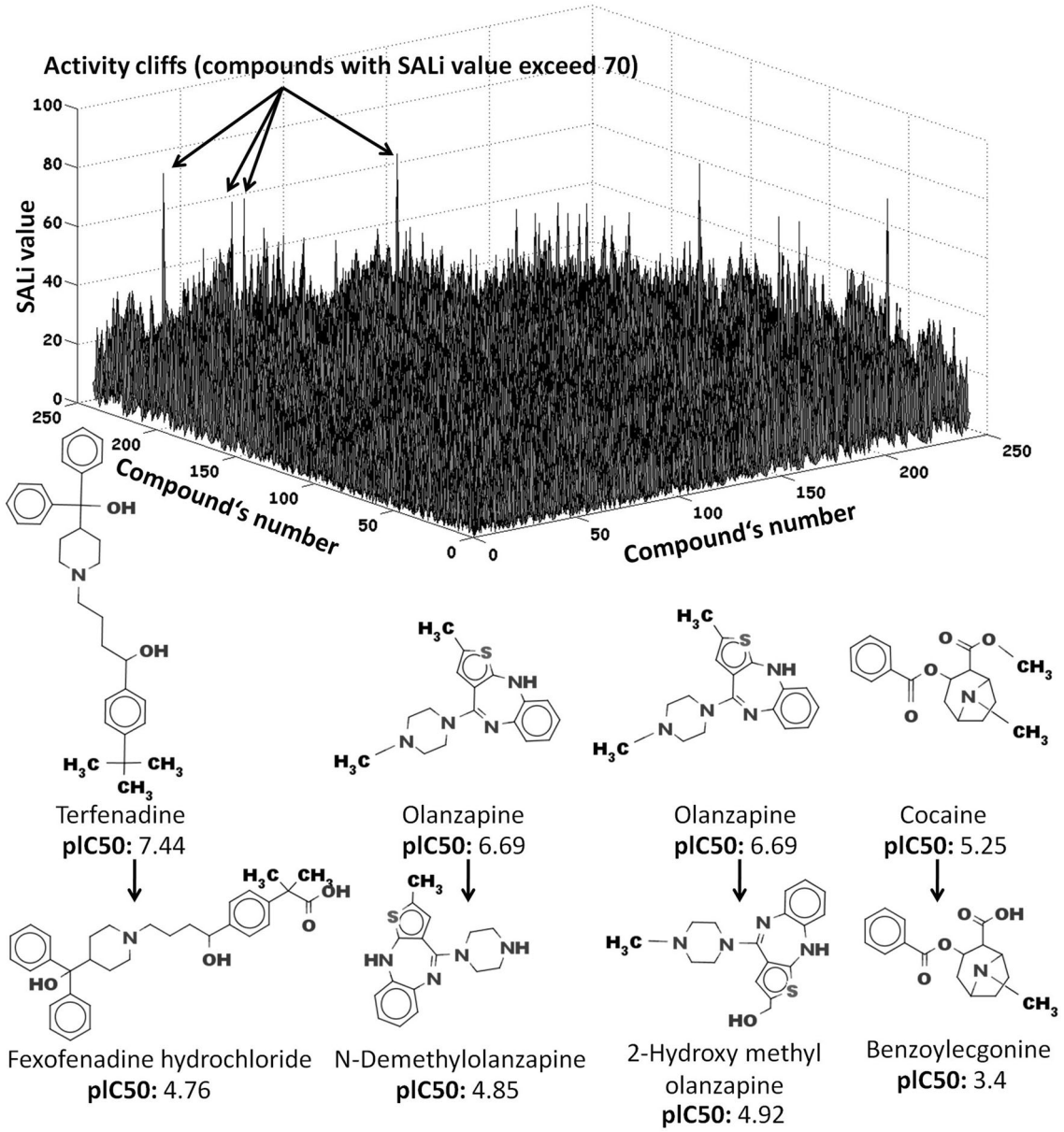


**Fig. 4** structure–activity Landscape Index (SALi) landscapes for Euclidean distances (before metric learning) (for HERG data set)

In each case, Balanced Accuracy values obtained using bootstrap aggregation exceeds the initial ones. This increase is demonstrated on the diagrams with the hatched areas on the bars and ranges from 0.04 up to 0.15. This observation corresponds with results published in the number of studies [46–48] where the use of bootstrap aggregation had been demonstrated as a means of improving the performance and robustness of the models. We decided to show these increase along with the original result as an interesting detail, but the main goal of bootstrap application was to determine the statistical significance of obtained results. Bootstrapping reduced to some extent differences between LMNN and MLMNN performances while preserved increasing BA difference between original and learned metric ($p$ value is not greater than 0.05).

The results were compared with the BA levels of the models developed with ASNN, one of the most known approaches of metric learning published in chemoinformatics. Figure 3 illustrates the results for models developed with and without applying PCA. One can see that in the second case there is almost no difference between models. For carcinogenicity data LMNN outperformed other methods by 0.04 after using the bootstrap aggregating procedure. Without bootstrap MLMNN performance for phospholipidosis was inferior to other methods. Using PCA some decrease of

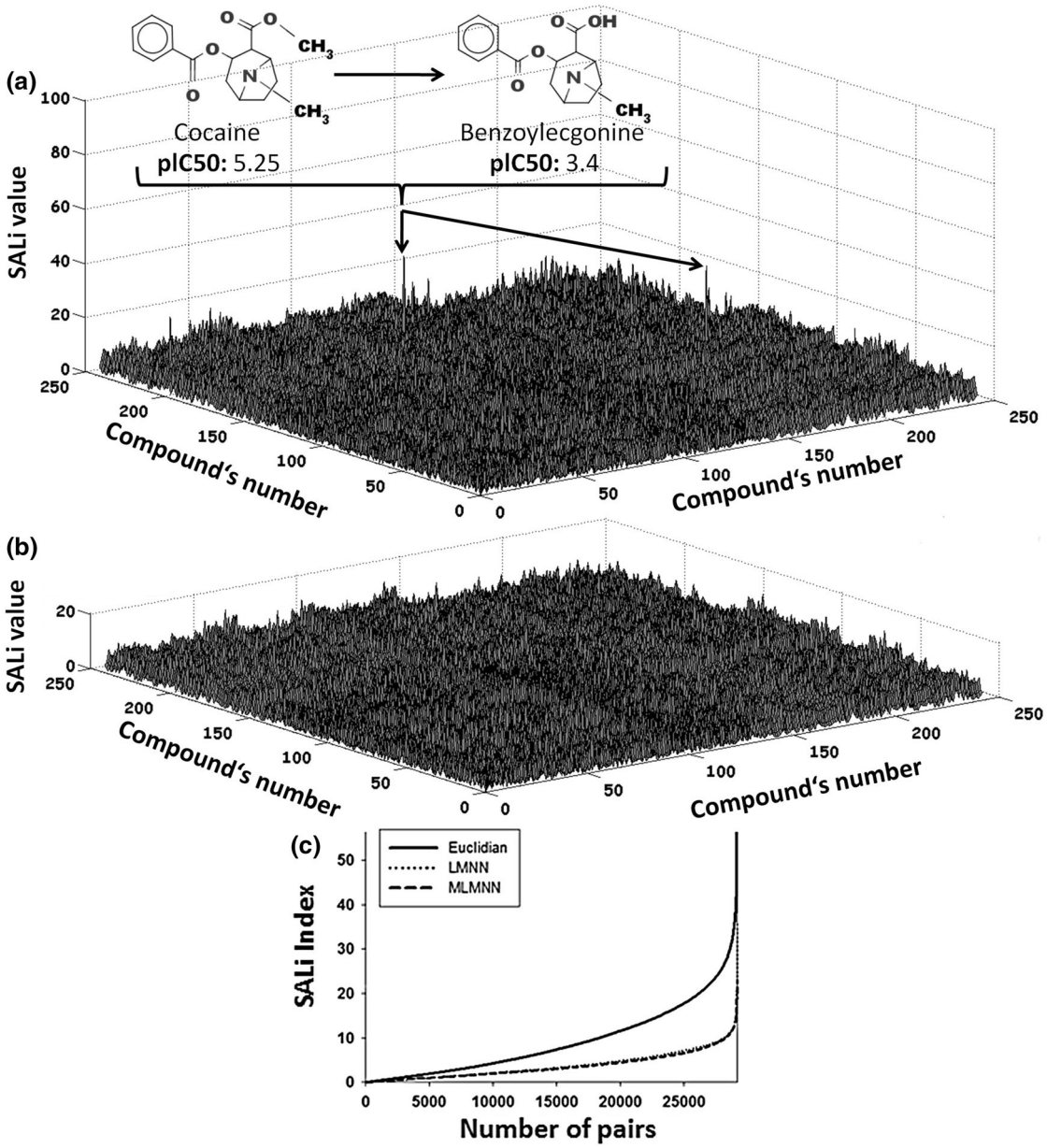


**Fig. 5** structure–activity Landscape Index (SALi) landscapes for learned metrics: **a** LMNN and **b** MLMNN (for HERG data set). The diagram **c** demonstrates the pairs of compounds ranked according to their SALi indices

performance of ASNN compared to LMNN and MLMNN was observed. This can be explained by the fact that parameters for PCA were optimized for *k*NN, but not for ASNN.

## Impact of parameters choice

There are several parameters that can be adjusted in LMNN procedure: (1) number of nearest neighbors $k$, (2) number of principal components *nfactors* and (3) weighting parameter μ (see Sect. "Metric learning"). To consider the impact of internal parameters choice on the predictive performance of classification models a grid search has been performed to adjust these parameters. The parameters search has been carried out using internal tuning set selected as a 1/3 part of each training set. Following parameters have been used to develop LMNN and MLMNN models (nfactors are given for the cases when PCA has been used as a pre-processing step): hERG $k = 3$; *nfactors* $= 82$; $\mu = 0.5$; phospholipdosis $k = 3$; *nfactors* $= 28$; $\mu = 0.5$; and carcinogenicity $k = 3$; *nfactors* $= 150$; $\mu = 0.5$. It was found that although the parameter $\mu$ can be adjusted by means of cross-validation procedure, according to our experience, the classification performance did not depend significantly on the values of $\mu$

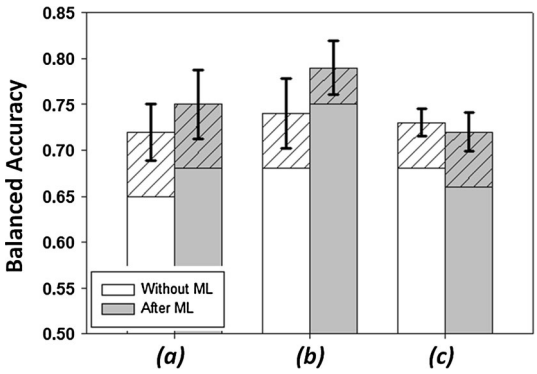


**Fig. 6** *Balanced Accuracy* of classification for *SVM* classifier for **a** HERG, **b** phospholipidosis and **c** carcinogenicity data sets. The hatched areas demonstrate the increase of BA after application of bootstrap procedure

and $k$ and the default values of $\mu = 0.5$ and k = 3 worked well. Metric learning approaches are sensitive to the number *of principal components* (nfactors) if PCA is used as preprocessing step which should be optimized in cross-validation.

## Structure–activity landscape indices (SALi)

The pair-wise Euclidean distances between all compounds were systematically computed for hERG data set (data set with measured activity values). For obtained distance matrices the structure–activity landscape indices (SALi) have been computed, where SALi is the ratio of the absolute value of the activity difference to the distance in the chemical space between pair of compounds. The same procedure has been carried out for the distances obtained after metric learning procedures (both LMNN and MLMNN). Finally, we obtain two matrices NxN, where N is number of the compounds in the data set. The visualization of obtained SALi landscapes is represented in Figs. 4 and 5. The application of PCA does not significantly change SALi landscape so the compounds that can be considered as activity cliffs are the same using PCA as a pre-processing step or not. In Figures the results without PCA application are given. One can see that the number of activity cliffs was reduced after metric learning procedure. The cases with SALi value exceeded 70 have been found only for initial descriptor space (Fig. 4). Four pairs of compounds were recognized: terfenadine and fexofenadine hydrochloride (pIC50 = 7.44 and 4.76, accordingly); olanzapine and *N* demethylolanzapine (pIC50 = 6.69 and 4.85, accordingly); olanzapine and 2-Hydroxy methyl olanzapine (pIC50 = 6.69 and 4.92, accordingly); cocaine and benzoylecgonine (pIC50 = 5.2 and 3.4, accordingly).

After metric learning procedure the landscapes have been changed. The number of activity cliffs after metric learning procedure has been decreased. In Fig. 5 only one pair of compounds that can be still considered as the activity cliffs in the new metrics remains (for LMNN-modified metrics). The analysis of the dataset showed that the compounds with structure similar to cocaine and benzoylecgonine are rare in the set. This lack of examples may have obstructed the proper metric learning by LMNN.

The diagram in Fig. 5c demonstrates the pairs of compounds ranked according to their SALi indexes. One can see that after metric learning the activity landscape became much smoother.

## Support vector machines and metric learning

In order to check if the learned metric with LMNN parameters optimized for *k*NN will work with other machine learning approaches as well we have used the linear transformation function L to transfer the metric in SVM (computational details are given in Sect. "Computational procedures"). In Fig. 6 one can see that the predictive performance of SVM models has been only insignificantly changed after metric learning procedure (results for external test set) that does not allow us to arrive at a conclusion about universality of proposed metric learning approach. The performance of obtained results concerned with parameter optimization of the LMNN that should be carried out with respect to a specific machine

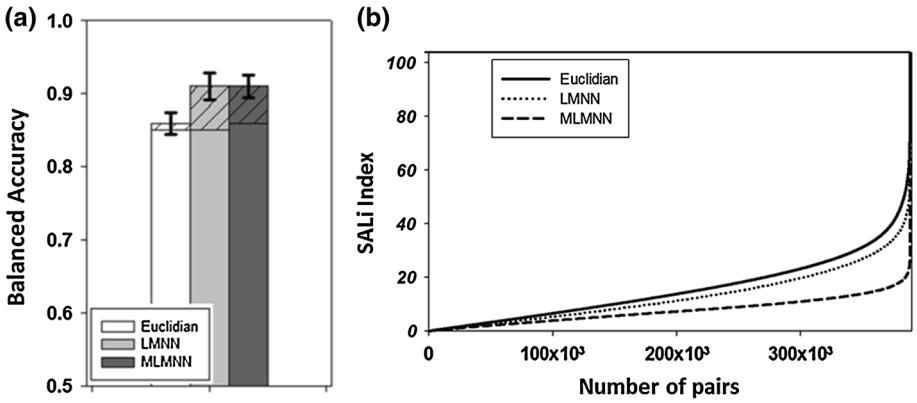


**Fig. 7** Distance Metric Learning without PCA for pre-processing: **a** *Balanced Accuracy* for the set of Xa inhibitors. The hatched areas demonstrate the increase of BA after application of bootstrap procedure. The diagram **b** demonstrates the pairs of compounds ranked according to their SALI indices

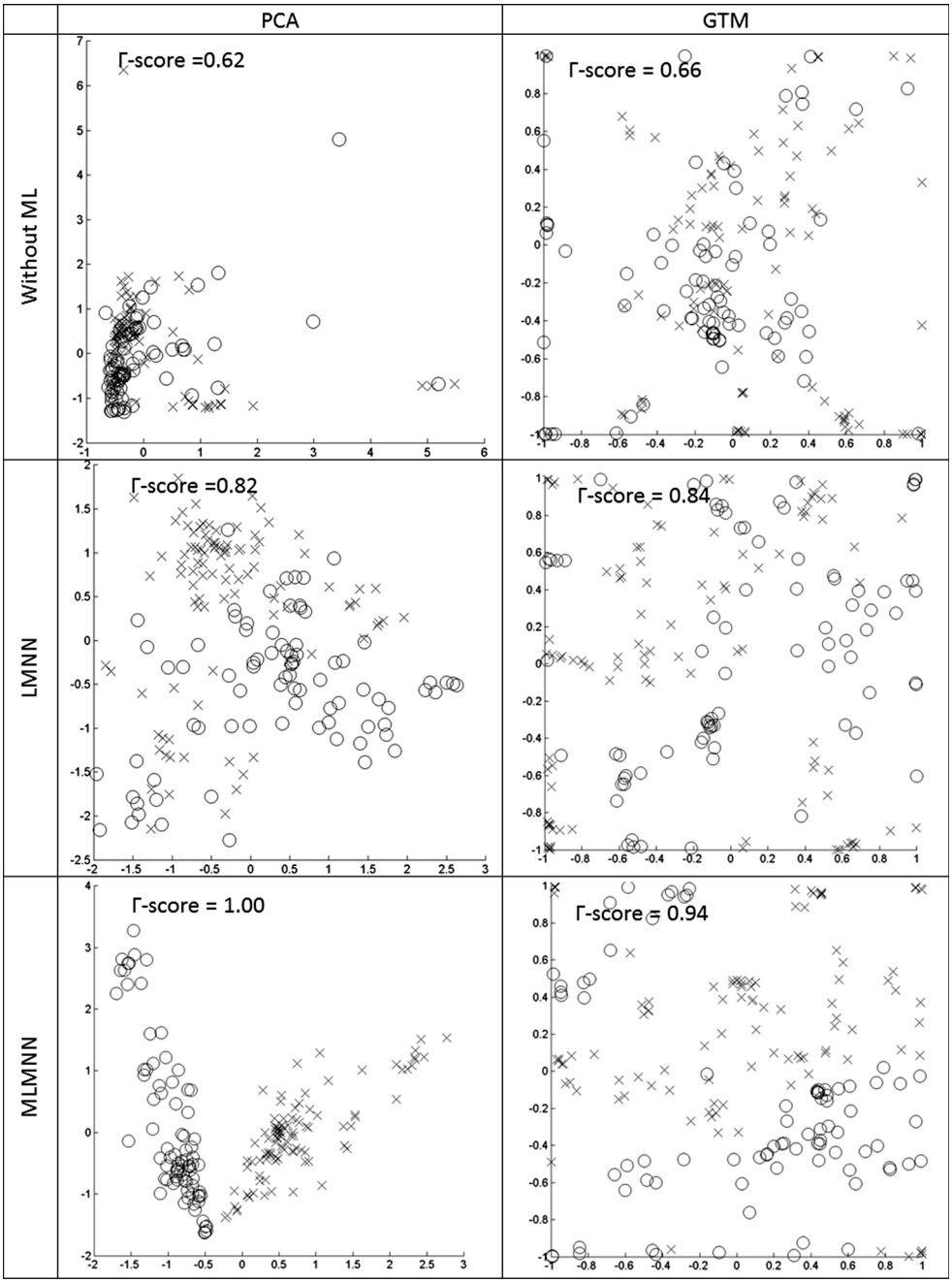


**Fig. 8** Impact of metric learning to data visualization performance: PCA and GTM visualization of phospholipidosis data set (here, x—active, o—inactive)

learning approach used for model development. These results conform to those presented in [49]. This may be attributed to the correspondence of optimal hyperplanes in the original and linearly transformed data spaces. In [50] authors demonstrated that SVM can be formulated as a metric learning problem that provides new insights to it and LMNN can be considered as learning a set of local SVM-like models in a quadratic space. At the same time, recently, number of studies [49–52] have been published that report using different metric learning approaches to support vector machines. The results of these studies are different, some of them indicate an improvement in the results performance while others claim the opposite observations. Perhaps, it can be data-dependent.

Before metric learning SVM and *k*NN demonstrated similar results. *k*NN outperforms SVM on hERG dataset (Balanced Accuracy 0.65 and 0.70 respectively) and SVM exceeds the results of *k*NN for carcinogenicity dataset (BA 0.68 and 0.62 respectively). After bootstrapping the latter difference was found insignificant (BA 0.73 and 0.72 respectively). The metric learning more affected the performance of *k*NN rather than of SVM. After it the values of

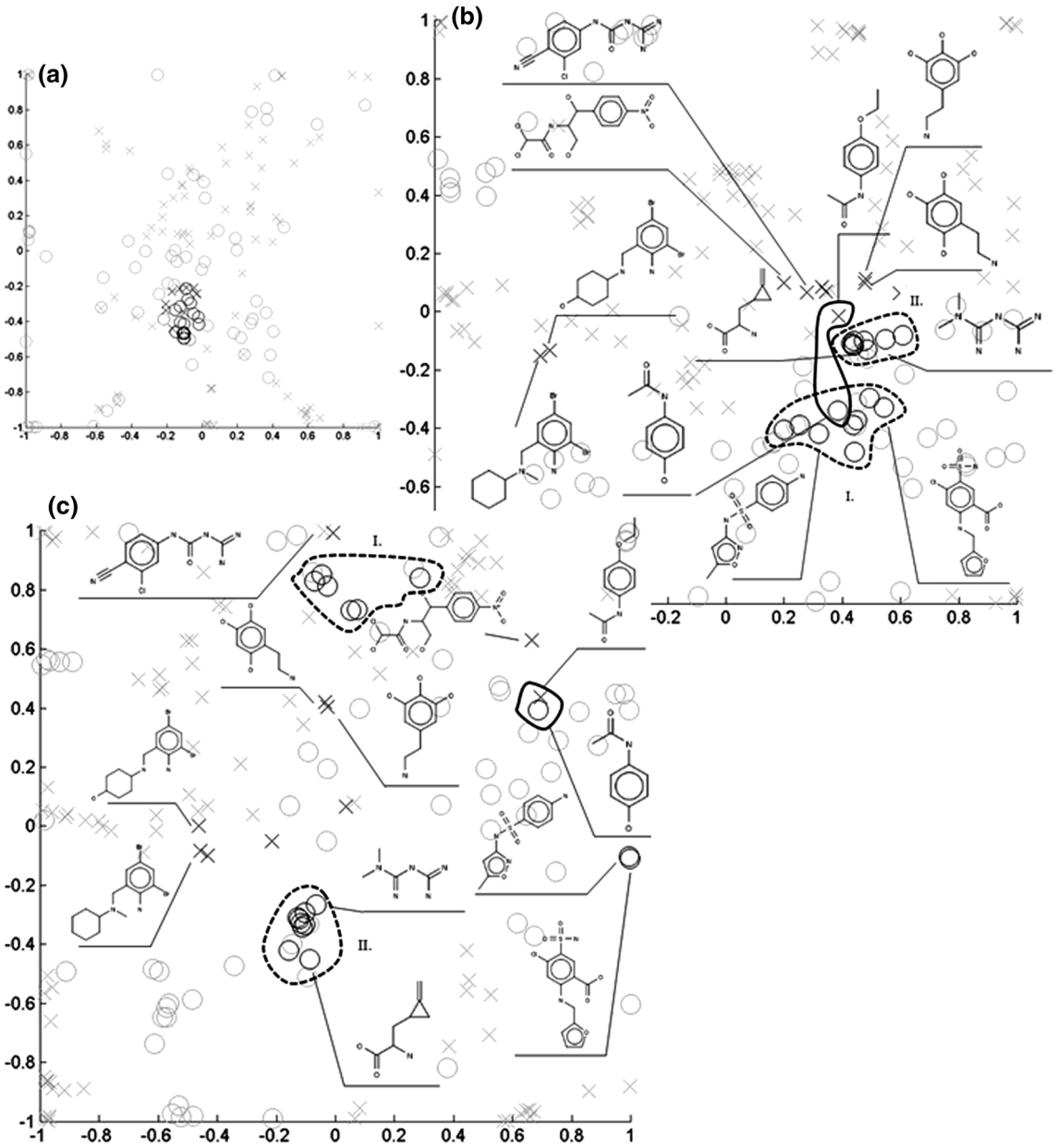


**Fig. 9** Analysis of visualization maps of phospholipidosis data set. On GTM visualization map obtained without metric learning (**a**) a small cluster of compounds have been chosen. These compounds have been highlighted after applying local (**b**) and global (**c**) metric learning (below). Structurally similar compounds that are encircled with *solid line* are related to different classes, but still mapped close to each other

Balanced Accuracy of all *k*NN models were larger than corresponding values for SVM, yet only the difference for hERG dataset can be found significant (BA 0.75 and 0.85 for SVM and *k*NN, respectively). One can come to a conclusion that *k*NN combined with a considered metric learning approaches can provide comparable or better results of those obtained with SVM especially taking into account basically easier and intuitively understandable model interpretation. Nevertheless it should be mentioned that LMNN approach that has been developed to be combined with *k*NN can be not so efficient in combination in SVM. Thus, recently number of alternative metric learning methods have been developed and adapted to use with SVM.

## Predictive performance for a "structure–activity" dataset

To see if the effect of metric learning can be observed for other objects, the brief additional analysis for the set of direct factor Xa inhibitors retrieved from [27] was

performed. The dataset was prepared as described in Sect. "Data and descriptors" The classification models were developed in the same manner as it had been done for other datasets, without PCA. The obtained results are demonstrated on the Fig. 7a. One can notice some increase in BA level from 0.86 to 0.91 for both LMNN and MLMNN, and though it appeared only after application of bootstrap aggregating procedure, its statistical significance is rather high ($p$ value less than 0.01).

An analysis of the structure–activity landscape has been performed in the similar wy to those of hERG data set [see Sect. "Structure–activity Landscape Indices (SALi)"]. Figure 7b represents the pairs of compounds ranked according to their SALi indices. One can see, a certain modifications of the activity landscape after metric learning procedure. The initial activity landscape contained 38 peaks for which SALi indices exceed defined threshold value. Only 6 observed peaks remained after global metric learning while there is no peaks exceeds defined threshold after multi-metric learning procedure.

### Metric learning and data visualization

In order to analyze the impact of metric learning to data visualization performance (for many cases, the classification and visualization goals do not coincide) we have used the linear transformation function L to transfer the learned metric in PCA and GTM (see Sect. "Data visualization techniques"). To this end, we have learned metrics (both local and global ones) for the whole data sets and used the modified descriptor spaces for the development of data visualization models. The example of the results (phospholipidosis dataset) for both data visualization approaches (PCA and GTM) are demonstrated in Fig. 8. For Euclidian metrics and for globally learned metric the nonlinear models (GTM) are characterized by better model performance irrespective of the data set has been used. The application of local metric learning approach leads to greater PCA results for all considered data sets. Metric learning (both local and global) significantly increase model performances in all cases but one (visualization carcinogenicity data set using PCA has no improvements after global metric learning).

GTM model for visualization of phospholipidosis data set has been analyzed (Fig. 9). We considered a cluster of overlapped compounds from both classes and checked how their location changed after metric learning. It was observed that local metric learning tends to keep identically labeled compounds tight while separating compounds from different classes despite of their structural similarities (see the compounds encircled with a solid line in Fig. 9). Global metric learning also keeps differently labeled compounds apart but distribute them along the map more dispersed, mapping compounds with similar structure together. Both local and global metric learning managed to divide the selected inactive compounds into two distinguished groups (labeled as (I) and (II) and encircled with dotted lines). Representatives of these two groups have significant differences in structure but without metric learning overlapped each other during visualization. Moreover considered groups obtained by applying MLMNN contain all the compounds of corresponding LMNN groups.

## Conclusions

This study concerns large margin nearest neighbors classifier and its multi-metric extension as the efficient approaches for metric learning which aimed to learn an appropriate distance/similarity function for considered case studies. In recent years, many studies in data mining and pattern recognition have demonstrated that a learned metric can significantly improve the performance in classification, clustering and retrieval tasks. The paper describes application of the metric learning approach to in silico assessment of chemical liabilities. Chemical liabilities, such as adverse effects and toxicity, play a significant role in 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. In this work, several aspects of metric learning procedure have been considered. First, the significant improvement in the performance of *kNN* classification models was shown as a result of carried out metric learning procedure for four data sets. Secondly, and here to our knowledge for the first time, the impact of metric learning on activity landscapes has been analyzed. The learned metric has been used in support vector machines. The obtained results allow one to conclude that predictive performance concerned with parameter optimization of the LMNN. Latter should be carried out with respect to a specific machine learning approach used for model development. Finally, the results of metric learning have been illustrated using linear and non-linear data visualization techniques in order to indicate how the change of metrics affected nearest neighbors relations and descriptor space. It was shown that metric learning improve the visualization performance of models irrespective of the chemography approach used.

**Acknowledgments** Authors thank Russian Foundation for Basic Research (Projects No. 11-03-00161 and 12-03-33086) for the support. NK acknowledges Dr. Igor Tetko for the provided ASNN software and Prof. Alexandre Varnek and Dr. Igor Baskin for their help and advice.

## References


1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214
2. van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2(3):192–204
3. Chopra S, Hadsell R, LeCun Y (2005) Learning a similiarty metric discriminatively, with application to face verification. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, San Diego
4. Domeniconi C, Gunopulos D, Peng J (2005) Large margin nearest neighbor classifiers. IEEE Trans Neural Netw 16(4):899–909
5. Goldberger J, Roweis S, Hinton G, Salakhutdinov R (2005) Neighbourhood components analysis. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems, vol 17. MIT Press, Cambidge, pp 513–520
6. Shalev-Shwartz S, Singer Y, Ng AY (2004) Online and batch learning of pseudo-metrics. In: Proceedings of the 21st International Conference on Machine Learning, Banff
7. Xing EP, Ng AY, Jordan MI, Russel S (2002) Distance metric learning, with application to clustering with side-information. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing systems, vol 14. MIT press, Cambridge
8. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. JMLR 10:207–244
9. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27
10. Horvath D, Barbosa F (2004) Neighborhood behavior—the relation between chemical similarity and property similarity. Curr Trends Med Chem 4:589–600
11. Horvath D, Jeandenans C (2003) Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces—a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. J Chem Inf Comput Sci 43(2):680–690
12. Horvath D, Koch C, Schneider G, Marcou G, Varnek A (2011) Local neighborhood behavior in a combinatorial library context. J Comput Aided Mol Des 25(3):237–252
13. Keefer CE, Kauffman GW, Gupta RR (2013) Interpretable, probability-based confidence metric for continuous quantitative structure–activity relationship models. J Chem Inf Model 53(2):368–383. doi:10.1021/ci300554t
14. McLellan MR, Ryan MD, Breneman CM Rank order entropy: why one metric is not enough. J Chem Inf Model 51(9):2302–2319. doi:10.1021/ci200170k
15. Skvortsova MI, Baskin II, Stankevich IV, Palyulin VA, Zefirov NS (1998) Molecular similarity. 1. Analytical description of the set of graph similarity measures. J Chem Inf Comput Sci 38(5):785–790
16. Tetko IV (2002) Neural network studies. 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42(3):717–728
17. Tetko IV (2002) Associative neural network. Neural Process Lett 16(2):187
18. Bajorath J (2012) Modeling of activity landscapes for drug discovery. Expert Opin Drug Discov 7(6):463–473
19. Guha R (2012) Exploring structure–activity data using the landscape paradigm. Wiley Interdiscip Rev: Comput Mol Sci 2(6):829–841. doi:10.1002/wcms.1087
20. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure–activity relationship analysis. J Med Chem 53(23):8209–8223. doi:10.1021/jm100933w
21. Iyer P, Stumpfe D, Vogt M, Bajorath J, Maggiora GM (2013) Activity landscapes, information theory, and structure—activity relationships. Mol Inf. doi:10.1002/minf.201200120
22. Guha R (2008) On the interpretation and interpretability of quantitative structure–activity relationship models. J Comput Aided Mol Des 22(12):857–871
23. Guha R, Van Drie JH (2008) Structure–Activity Landscape Index: identifying and quantifying activity cliffs. J Chem Inf Model 48(3):646–658. doi:10.1021/ci7004093
24. Nisius B, Goller AH, Bajorath J (2009) Combining cluster analysis, feature selection and multiple support vector machine models for the identification of human ether-a-go–go related gene channel blocking compounds. Chem Biol Drug Des 73(1):17–25
25. DSSTox database. http://wwwepagov/ncct/dsstox/
26. Lowe R, Mussa HY, Nigsch F, Glen RC, Mitchell JB (2012) Predicting the mechanism of phospholipidosis. J Chemoinf 4:2
27. Wawer M, Jr Bajorath (2010) Similarity–potency trees: a method to search for SAR information in compound data sets and derive SAR rules. J Chem Inf Model 50(8):1395–1409
28. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29(6–7):476–488. doi:10.1002/minf.201000061
29. Chemaxon Standardizer. http://wwwchemaxoncom/library/scientific-presentations/standardizer/
30. Instant JChem www.chemaxon.com/products/instant-jchem/
31. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G (2008) ISIDA—platform for virtual screening based on fragment and pharmacophoric descriptors. Curr Comput Aided Drug Des 4(3):191–198
32. Swamy MNS, Thulasiraman K (1981) Graphs, networks, and algorithms. Wiley, New York
33. Boyd SP, Vandenberghe L (2004) Convex optimization. Cambridge university press, Cambridge
34. Jolliffe IT (2002) Principal component analysis. Springer series in statistics, vol XXIX, 2nd edn, Springer: NY
35. Bishop CM, Svensen M (1998) GTM: the generative topographic mapping. Neural Comput 10(1):215–234
36. Bishop CM, Svensen M, Williams CLI (1997) GTM: A principled alternative to the self-organizing map. Tech Rep Neural Comput Res Group
37. Maniyar DM, Nabney IT, Williams BS, Sewing A (2006) Data visualization during the early stages of drug discovery. J Chem Inf Model 46(4):1806–1818. doi:10.1021/ci050471a
38. Owen JR, Nabney I, Medina-Franco JL, Lopez-Vallejo F (2011) Visualization of molecular fingerprints. J Chem Inf Model 51:1552–1563
39. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A (2012) Generative topographic maps (GTM): universal tool for data visualization, structure–activity modeling and database comparison. Mol Inf 31(3–4):301–312
40. Kireeva N, Kuznetsov SL, Bykov AA, Yu Tsivadze A (2013) Towards in silico identification of the human ether-a-go–go-related gene channel blockers: discriminative vs. generative classification models. SAR QSAR Environ Res 24(2):103–117. doi:10.1080/1062936x.2012.742135
41. Kireeva N, Kuznetsov SL, Tsivadze AY (2012) Toward navigating chemical space of ionic liquids: prediction of melting points using generative topographic maps. Ind Eng Chem Res 51(44):14337–14343. doi:10.1021/ie3021895
42. Hasegawa K, Funatsu K Prediction of protein–protein interaction pocket using L-shaped PLS approach and its visualizations by generative topographic mapping. Mol Inf. doi:10.1002/minf.201300137
43. http://www.cse.wustl.edu/∼kilian/code/code.html
44. Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for

performance evaluation. In: AAAI Workshop—technical report, pp 24–29
45. Chang C-C, Lin C-J (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
46. Dias JG, Vermunt JK (2008) A bootstrap-based aggregate classifier for model-based clustering. Comput Stat 23(4):643–659
47. Barutçuoğlu Z, Alpaydın E (2003) A comparison of model aggregation methods for regression. In: Artificial neural networks and neural information processing—ICANN/ICONIP. Springer, pp 76–83
48. Varnek A, Baskin I (2012) Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model 52(6):1413–1437
49. Xu Z, Weinberger KQ, Chapelle O (2012) Distance metric learning for kernel machines. arXiv preprint arXiv:12083422
50. Do H, Kalousis A, Wang J, Woznica A (2012) A metric learning perspective of SVM: on the relation of LMNN and SVM. In: International Conference on Artificial Intelligence and Statistics, pp 308–317
51. Liu Y, Caselles V (2011) Improved support vector machines with distance metric learning. In: Advances concepts for intelligent vision systems. Springer, pp 82–91
52. Zhu X, Gong P, Zhao Z, Zhang C (2012) Learning similarity metric with SVM. In: Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, pp 1–8