Received: 14 March 2015, Revised: 4 December 2015, Accepted: 11 December 2015, Published online in Wiley Online Library: 9 February 2016

(wileyonlinelibrary.com) DOI: 10.1002/cem.2776

Activity prediction and identification of mis-annotated chemical compounds using extreme descriptors

Petro Borysov[a,b]*, Jan Hannig[b], J. S. Marron[b], Eugene Muratov[c,d], Denis Fourches[c] and Alexander Tropsha[c]

Data pre-processing that includes removal of descriptors with low variance is a standard first step in quantitative structure–activity relationship modeling. In this paper, we study low-variance descriptors and show that some of them contain significant amounts of useful information. In particular, we define the notion of extreme descriptors (those variables that have the same value for almost all compounds and only a few values that are different from the common median). We show that extreme descriptors can be helpful for activity prediction in a standard binary classification setting. Moreover, we demonstrate using two case studies (M_2 muscarinic receptors and skin sensitization) that extreme descriptors can be used for the identification of possibly mislabeled compounds. Because of these previously unknown, but important, properties, extreme descriptors should be considered in quantitative structure–activity relationship modeling studies. Copyright © 2016 John Wiley & Sons, Ltd.

Additional information may be found in the online version of this article at the publisher's web site

Keywords: extreme descriptors; low-variance descriptors; mislabeling; QSAR; Prediction

1. INTRODUCTION

Data quality is critical for development of robust and predictive quantitative structure–activity relationship (QSAR) models [1,2]. Young *et al.* [2] showed that even a reasonably small (up to 4%) fraction of corrupted data can lead to significant decrease of model quality. This conclusion becomes especially important in light of the studies demonstrating that on average, there are two errors per each medicinal chemistry publication with an overall error rate for compounds in primary sources used to compile the WOMBAT database as high as 8% [3,4]. One of the sources of errors in chemical databases and data sets is a mislabeling of activity of investigated compounds. There could be numerous reasons for such artifacts, for example, erroneous transition from a publication to a database and shifting the string of labels toward the list of corresponding compounds.

The choice of structural descriptors and modeling techniques also has an influence on the quality of QSAR models. The recent study by Zhu *et al.* [5] demonstrated that the choice of descriptors are more important than the model optimization techniques. Nowadays, thousands of descriptors can be generated by different software packages for every compound in the data set. However, many of the descriptors do not contain any useful modeling information and just reduce the quality and interpretability of models based on them [6]. Therefore, the standard first step in QSAR modeling is data pre-processing, which generally includes removal of descriptors with low variance. However, the selection of a low-variance threshold is subjective and often is not data set specific. Also, it is possible that descriptors with variance smaller than selected threshold can still contain some useful information. In this paper, we will show why descriptors that have the same value for almost all compounds and only a small fraction of values that are different from the common median should be used in the modeling process to enhance the performance of standard methods. Furthermore, we propose a measure of prediction confidence based on extreme descriptors (EDs) and only consider prediction with high confidence in our analysis.

There are several methods that can reduce the high dimension of the data to a lower dimension, for example, principle component analysis [7] or partial least squares [8], so that classical multivariate techniques can be applied. In recent years, there was a lot of attention paid to regularized sparse regression methods,

* *Correspondence to: Petro Borysov, SAS Institute, Cary, NC 27513, USA*
E-mail: pborysov@gmail.com

a *P. Borysov*
SAS Institute, Cary, NC 27513, USA

b *P. Borysov, J. Hannig, J. S. Marron*
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA

c *E. Muratov, D. Fourches, A. Tropsha*
Department of Chemical Biology and Medicinal Chemistry, University of North Carolina, Chapel Hill, NC 27599, USA

d *E. Muratov*
A.V. Bogatsky Physical–Chemical Institute National Academy of Sciences of Ukraine, Odessa, 65080, Ukraine

for example, least absolute shrinkage and regression operator (LASSO) [9]. Sparse methods were also developed for classification problems, for example, sparse support vector machine (SVM) [10]. A weakness of the first sparse methods was in their inability to find more than one, often inadequate, representative of a set of descriptors that work as a group. This drawback has been addressed by methods such as group LASSO [11]. It was designed with a goal of finding important explanatory factors that may be represented by a group of variables. Even though all these methods are very powerful, usually, they are applied to the preprocessed data where many descriptors, especially with low variance, are removed. In this paper, we introduce the term *extreme descriptor* to define descriptor that has almost all values the same and all different values in just one class (see panels B–D of Figure 1). EDs are data set dependent. That is, the same descriptor could have extreme values for one data set and regular for another one. Expectedly, EDs have small overall variance, but we will show that they can be used not only for prediction but also for identification of mislabeled compounds. Although the majority of EDs are fragment descriptors, certain integral descriptors could also match our definition.

The situation is different for the quality of labels because the errors are completely contained in the training set and are not extended to new data points. Brodley and Friedl [12] identified several sources of labeling errors, such as subjectivity, data entry, and inadequacy of the information, used to label each observation . These errors potentially lead to contradictory labels, where the same observations appear more than once and belong to different classes. They also may lead to misclassifications, where observations are assigned to incorrect classes. Thus, the identification and removal of the misclassified observations in many situations substantially improve the performance of the model [13].

The problem of misclassification was addressed previously by researchers in several areas, especially in genetics and medicine. For example, Zhang *et al.* [14] proposed the procedure for handling potential mislabeling among training samples based on gene expression data in human breast cancer study. Joseph *et al.* [15] analyzed and validated the reclassification of several subjects that were misdiagnosed with Alzheimer's disease. Gamberger *et al.* [16] studied mislabeling in early diagnosis of rheumatic diseases. Brodley and Friedl [12] investigated the mislabeling problem in areas of automated land cover mapping and credit approval. Many of the proposed methods rely on the idea of outlier removal in regression analysis. Wilson and Martinez [17] used various versions of the k-nearest neighbor classifier as a filter to identify and eliminate suspect observations. Brodley and Friedl [13] applied an ensemble of classifiers with several voting strategies. Zeng and Martinez [18] proposed an *automatic data enhancement* approach based on the mechanisms of neural networks to correct mislabeled data points. Teng [19] developed a two-stage decision tree classifier designed to correct noise both in labels and in variables. Gamberger *et al.* [16] proposed a noise detection and elimination method based on the minimum description principle. In the area of QSAR, Fourches *et al.* [20] showed that consensus models can be used to flag and correct bioactivity annotation for certain compounds in a data set. Fourches *et al.* [1] demonstrated the success of the consensus models approach on Ames mutagenicity data set [21,22].

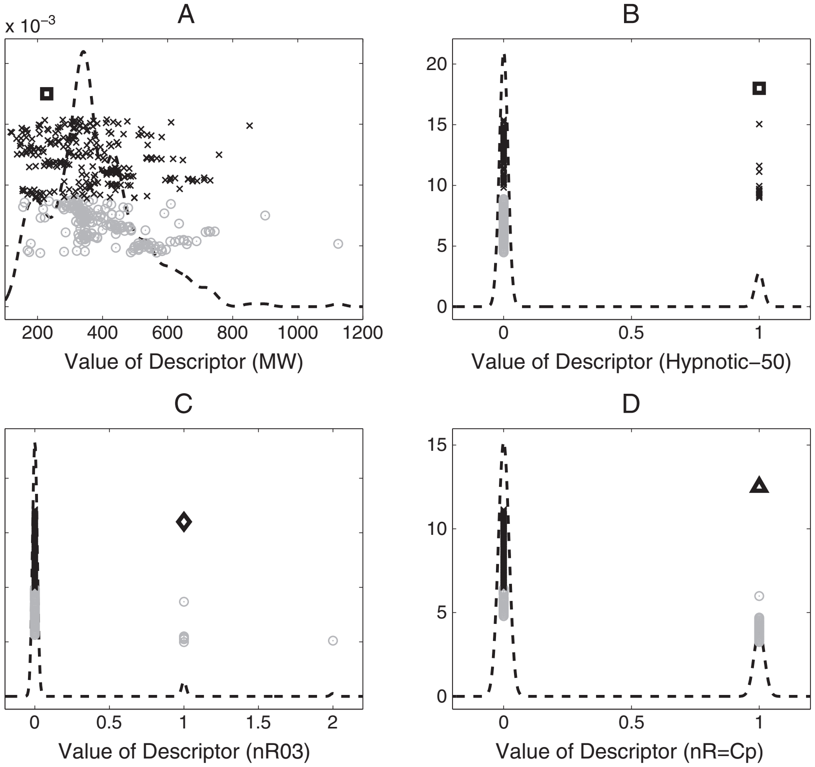

Figure 1. Distributions of regular (A) and extreme (B–D) descriptors on their original scale from the M_2 muscarinic receptors data. Gray circles and black crosses represent 179 active and 265 inactive compounds, respectively. The black dashed curve is a smooth histogram. Test compounds are shown as black square, diamond, and triangle. EDs have zero variance in one of the classes. Test compounds were predicted using the EDs as inactive (B) and as actives (C and D).

In this data set, 31 compounds were erroneously annotated in the original database and successfully corrected. Recently, the same approach was used for the identification of two mislabeled compounds in reduced Local Lymph Node Assay (rLLNA) skin sensitization data set [23].

Descriptor selection and preprocessing should be performed with caution. Quinlan [24] showed that attempts at removing noise from the variables decrease the predictive performance of the classifier when the same noise level is present in the data to be predicted. The goals of this study are as follows: (i) to introduce the concept of EDs; (ii) to prove using case studies that those descriptors are useful for improving the prediction of new compounds; and, the most important, (iii) to demonstrate that they can be used for identification of mislabeled compounds.

2. DATA

In this paper, we present the analysis of two chemical data sets using EDs. The first data set used in this study was compiled by Alves *et al.* [23] from the Interagency Coordinating Committee on the Validation of Alternative Methods report [25]. Each chemical was designated as sensitizer/non-sensitizer according to the value of its effective concentration (EC_3). The compounds were tested in multiple settings to achieve optimal solubility and skin penetration. If for any compound, conflicting sensitization potentials were found, then such compounds were removed from the data set. This process resulted in 381 (253 sensitizers and 128 non-sensitizers) unique data points that were further employed for modeling. This data set was unbalanced, and in order to avoid QSAR models with biased predictivity, we balanced it before starting the modeling. Instead of randomly removing a certain proportion of sensitizers from the data set, we performed a similarity search relying on non-sensitizers as a starting point to search the active pool for structurally similar compounds. This similarity-based selection procedure was carried out by the methods of data analysis module of the HiT QSAR software [26] in two stages: (i) generate the matrix of Euclidean distances in the chemical space between all the pairs of compounds, and then (ii) choose 64 sensitizers with the smallest Euclidean distance to the nearest non-sensitizer. The final data set consisted of 262 compounds: 134 sensitizers and 128 non-sensitizers.

The second data set was compiled with compounds tested against M_2 muscarinic receptors located in the heart. Their function of M_2 muscarinic receptors is to slow down the heart rate to normal rhythm after stimulation by the sympathetic nervous system. These receptors also reduce contractile forces of the atrial cardiac muscle; however, they have no effect on the contractile forces of the ventricular muscle. The data set consists of 444 compounds: 179 active and 265 inactive.

Before descriptor generation step, compounds from both data sets were carefully curated following the work flow described by Fourches *et al.* [1]. In the next step, the following types of descriptors were generated for standardized chemical structures using Dragon software (v.5.5; Talete SRL, Milan, Italy): 0D constitutional (atom and group counts), 1D functional groups, 1D atom-centered fragments, 2D topological descriptors, 2D walk and path counts, 2D autocorrelations, 2D connectivity indices, 2D information indices, 2D topological charge indices, 2D eigenvalue-based indices, 2D topological descriptors, 2D edge-adjacency indices, 2D burden eigenvalues, 2D binary fingerprints, 2D frequency fingerprints, and molecular properties. Detailed discussion of these descriptors can be found in Todeschini and Consonni [27].

3. IMPROVING PREDICTION ACCURACY OF NEW COMPOUNDS

This section describes the proposed method for identification of a set of compounds that are possibly assigned a wrong label. To predict the labels of chemical compounds, the proposed method uses EDs. First, we will explain how the predictions are made for each compound and then how possibly mislabeled compounds are identified.

Recall from Section 1 that we define EDs to have almost all values the same and a small fraction of different values in just one class. The behavior of regular descriptors is illustrated in panel A of Figure 1, and three special cases of EDs (with zero variance in one of the classes) are shown in panels B, C, and D. Most of the EDs are fragment counts (panels C and D); however, integral parameters are also represented (panel B).

Molecular weight is shown on Figure 1(A) as an example of regular descriptor, which is typically retained for analysis by many researchers. However, as seen in Figure 1(A), molecular weight by itself contains very little information for classification of the compound shown as a black square. The distribution of the descriptor (Hypnotic-50), which takes the value of 0 for all active compounds (gray circles) and almost all inactive compounds (black crosses) except a few inactives, for which Hypnotic-50 is equal to 1, is shown on Figure 1(B). Hypnotic-50 belongs to Ghose–Viswanadhan–Wendoloski drug-like indexes and provides an evaluation of “drug likeliness.” If Hypnotic-50 is equal to 1, it means that lypophilicity, molecular weight, and the total number of atoms in a given compound are within the range of corresponding properties covering 50% of drugs with hypnotic effect. The variance of this descriptor for the active compounds is zero. If a new compound has a value near 1, then this descriptor provides evidence that the new compound is inactive (black). Otherwise, this descriptor has no useful information. For example, consider the new compound shown as the black square in Figure 1(B). Because the value of the descriptor of this new compound is taken on only by inactive class members, it appears to be inactive. In this case, there are 12 inactive compounds that have a value of 1 for Hypnotic-50 descriptor and thus support this prediction. Clearly, the larger the number of compounds that support such a prediction, the higher the confidence that the prediction is correct. The reported label of the new compound (black square) is active, but EDs as well as a classification rule (described later in this section) assign this compound to the inactive class. Further verification showed that the test compound is mislabeled. Figure 1(C) is similar to Figure 1(B), except now descriptor (nR03 – number of 3-membered rings, e.g., aziridine or oxirane) has three different values for compounds from the active class. In this case, the predicted label of a new compound (black diamond) is active (same as the original label), while the classification rule assigns it to the inactive class. Figure 1(D) illustrates a prediction of another compound (black triangle) with the descriptor (nR = Cp – number of terminal primary carbons with sp^2-hybridization) that has zero variance within the inactive class. In this case, the predicted label of the new compound (originally labeled active) is taken to be active, and the number of compounds that support the label assignment is 19.

As mentioned earlier, EDs have very small variance and often removed from the analysis during data preprocessing steps. Additionally, the vast majority of EDs are not correlated with themselves and other descriptors, but in rare cases, the correlations are close to one. This phenomenon is attributed to the fact that for a specific ED, most of the compounds share the same value. Because most of EDs are fragment counts, sets of compounds that share the same values are different across descriptors. Hence, the correlation coefficient is near zero. In cases when sets of compounds that share the same values have significant overlap, the correlation is close to one. Furthermore, individual EDs have relatively small impact on dimension reduction methods. For example, the size of the average loading among EDs is only around 5% of the size of the average loading across the remaining descriptors. Therefore, we believe that EDs should be used collectively (described later) and, if possible, in combination with other methods.

In the first step of the proposed method, we divide the data into training and test sets. Then we identify EDs using only the training set compounds. The information about the predictions from all EDs identified in the training set is combined and used to assign a class label to some compounds in the test set. If the predicted class label for a particular compound does not match the original label and if the number of compounds that support the prediction is greater than a specified threshold, then this compound is a candidate for being misclassified. Detailed explanation about how the predictions from the individual EDs can be combined is provided in Label prediction using extreme descriptors section (Appendix). The estimation of the threshold is described in Section 3.2.

3.1. Extreme descriptors for data with mislabeling

The definition of EDs states that their values are identical for almost all compounds in a data set and only a small fraction of compounds from one of the classes have a value that is different from the common median. This definition can be extended to the case where compounds that have descriptor values different from the common median are not required to be from the same activity class.

Suppose that p is the percentage of the compounds that are mislabeled. According to Olah *et al.* [4], it is reasonable to assume that this proportion can be as high as 8% . We define a more general version of EDs as predictors that have one value shared by most of the compounds. The remaining values have almost all compounds in one class, with at most p% in the other class. This definition allows us to use these descriptors and also handle situations where the training data set contains potentially mislabeled compounds. Similar to Section 3, the prediction by each generalized ED will be supported by the set of compounds that belong to both classes. In such cases, the label is assigned by the majority vote. The existence of the majority is guaranteed by the selection process of EDs described in Appendix.

Note that EDs described earlier and illustrated in Figure 1(B)–(D) are special cases of generalized EDs with 0% of mislabeling. In the rest of this paper, we will use the term *extreme descriptor* to denote the general version that allows for mislabeling and will specify the mislabeling percentage when necessary.

3.2. Cutoff and label significance

This section describes the estimation of the cutoff on the number of compounds that support predictions by EDs. If the number of compounds that disqualify the current label is large enough, it is reasonable to suspect that there is a mismatch between the label and the chemical structure represented by the set of descriptors.

To illustrate the main idea of the cutoff estimation, without loss of generality, we only consider "true" inactive compounds that potentially could be mislabeled. The estimation for active compounds is carried out similarly.

Suppose that θ is the population proportion of inactive compounds that are predicted as active by EDs, that is, the probability that the proposed method will make incorrect prediction. For every compound in the test set that has EDs identified in the training set, we can construct a confidence interval for θ. The confidence interval will depend on two parameters, n and k. The first parameter n represents the total number of compounds in the training set that support prediction of any label across all EDs present in the test compound. The second parameter represents the number of compounds $k \in [0, n]$ in the training set that support only the active label. For example, compound 5 in Figure A1 has only two EDs (horizontal bars) and $k = 2$ active compounds from the training set that support a prediction of active label (two white dots on the gray horizontal bar) by EDs. At the same time, compound 5 has $n = 3$ compounds from the training set that support both labels (two active compounds that support an active label and one inactive compound that supports an inactive label). Similarly, for compound 1, $k = 0$, and $n = 59$. Using observed n and k, one can construct a $1 - \alpha$ confidence interval $(0, u)$ for the proportion θ. The upper bound u is the upper confidence limit for the original label. Naturally, it is also possible to find the number of compounds n required to guarantee that the misclassification probability is below the selected upper bound u with confidence level $1 - \alpha$.

Note that the described probability model only allows estimation of the confidence in the original label for those compounds that do not have contradicting predictions by EDs. Compounds that have very few descriptors with opposite predictions should be examined on a case by case basis.

The confidence interval for the parameter θ can be constructed using ideas of *pivoting a discrete pdf* [28] or generalized fiducial inference [29–32]. For example, the upper bound of the confidence interval for $k = 0$ can be obtained from the following equation:

$$1 - \alpha = \frac{1}{2} + \frac{1}{2}\int_0^u (1 - t)^{n-1} dt$$

which yields

$$u = 1 - (2\alpha)^{1/n} \tag{1}$$

This equation can be also solved for n to obtain the number of compounds n required to achieve upper bound u with confidence $1 - \alpha$. For example, to achieve a 10% upper bound on the probability that EDs made incorrect prediction with 95% confidence, it is required to observe $n = 22$. For different values of k, the relationship between parameters can be derived in a similar manner. Please see Appendix for details.

To estimate the upper bound u on the confidence of the original label using Equation (1), it is necessary to count only unique compounds observed across all EDs. But it is very possible that the same compound from the training set will support predictions made by several EDs. If this happens for many EDs, it means that the test compound is similar (has the same values of EDs) to the compounds in the training set. Therefore, such situations also should be carefully considered.

Table I. Analysis of prediction accuracy with and without EDs

Data	Description of prediction method	Active	Inactive	Sensitivity	Specificity	Accuracy	MCC
M_2 musc. rec.	DWD with all descriptors	179	265	81.01	66.04	72.07	0.46
	DWD without ED	179	265	80.45	66.42	72.07	0.46
	ED-improved DWD	179	265	82.68	67.17	73.42	0.49
	Accuracy bound	179	265	86.03	67.17	74.77	0.52
	ED only	33	44	81.82	100.0	92.21	0.85
	DWD for compounds with ED	33	44	72.73	93.18	84.42	0.68
Skin sens.	DWD with all descriptors	134	128	84.33	67.97	76.34	0.53
	DWD without ED	134	128	83.58	68.75	76.34	0.53
	ED-improved DWD	134	128	82.09	70.31	76.34	0.53
	Accuracy bound	134	128	91.04	70.31	80.92	0.63
	ED only	14	46	0.00	95.65	73.33	−0.11
	DWD for compounds with ED	14	46	21.43	89.13	73.33	0.13

Among other things shows maximum potential prediction improvement for inactive compounds by ED improved DWD classifier.

EDs, extreme descriptors; MCC, Matthews correlation coefficient; DWD, distance-weighted discrimination; musc. rec., muscarinic receptor; sens., sensitization.

4. ANALYSIS OF CHEMICAL DATA

In this section, we present the results of the analysis of two chemical data sets (included as supporting information) using EDs. In order to test the class label of all compounds, we use an external cross validation with every data set. There are several types of external cross validation. In general external *N*-fold cross validation, the data are randomly separated into *N* parts. One of those parts is used for blind prediction external test set, and the remaining $N - 1$ parts are used for model development. In QSAR, it is common to use fivefold external cross validation, while in the misclassification literature, the leave-one-out type ($N =$ sample size) is often used. In this paper, we adopt the idea of leave-one-out external cross validation with the proposed classifier based on the EDs, where in each fold, one compound is set aside. The collection of EDs as well as the classification model was developed based on the remaining data. Every compound is classified using EDs and the distance-weighted discrimination (DWD) classifier. For each compound that was classified using EDs, the original label confidence *u* was computed as described in the previous section. The predicted label was compared with the originally assigned class label. If for a particular compound the predicted label did not match the original label and its label upper confidence limit *u* was low, then those compounds were selected to be checked for mislabeling.

In this paper, we consider two versions of EDs: with no mislabeling (these descriptors have zero variance in one of the classes) and with possible mislabeling (when required, we will use 8%). In addition, during the preprocessing step, we did not remove descriptors with low variance. Only descriptors with zero total variance and all but one identical descriptor were removed from the data set.

4.1. Activity prediction with extreme descriptors

Depending on the objective, the analysis can be carried out in two opposite directions. In case of the prediction objective, the goal is to develop a model that is able to accurately predict the property of a new compound. One of the main assumptions here is that all labels in the training and validation data sets are correct. The accuracy of the model is estimated by the proportion of the predicted labels that match the original label for compounds in the validation data set. In case of identification of mislabeled compounds, the labels in the validation data set do not have to be correct. In this setting, if the predicted label is different from the original label in the validation data set, the compound becomes a suspect for being mislabeled.

In this section, we demonstrate how EDs can enhance standard classifiers from the prediction objective point of view. Table I shows the various aspects of the ED classifier's impact on the DWD classifier on M_2 muscarinic receptors and skin sensitization data sets. The prediction performance is estimated using sensitivity, specificity, accuracy, and Matthews correlation coefficient [33] (MCC) measures.

It is possible to see that simply removing EDs has a small effect on sensitivity and specificity, but not on the overall accuracy of the model for both data sets. Next, we combine predictions by EDs and DWD (see ED-improved DWD in Table I) in the following way. For all compounds that can be predicted with EDs, the prediction confidence is calculated using Equation (1). If the prediction confidence is large (at least 70 %), then the predicted label is based on the EDs alone; otherwise, it is based only on the DWD classifier. Addition of prediction by EDs improved the accuracy of the model (including sensitivity, specificity, and MCC) on the M_2 muscarinic receptors data. In the case of skin sensitization data set, the improvement was only in specificity at some cost of decline in sensitivity. Overall, accuracy and MCC remained the same.

Because the number of compounds that contain EDs is relatively small (around 20% of all compounds), the potential increase in accuracy cannot be very large. Even if the DWD classifier made correct prediction on all compounds with EDs, the potential increase would be about 20%. The accuracy bound for ED-improved DWD prediction method in Table I shows the highest possible accuracy assuming all compounds with EDs are correctly predicted. The ED-improved DWD classifier achieved maximum

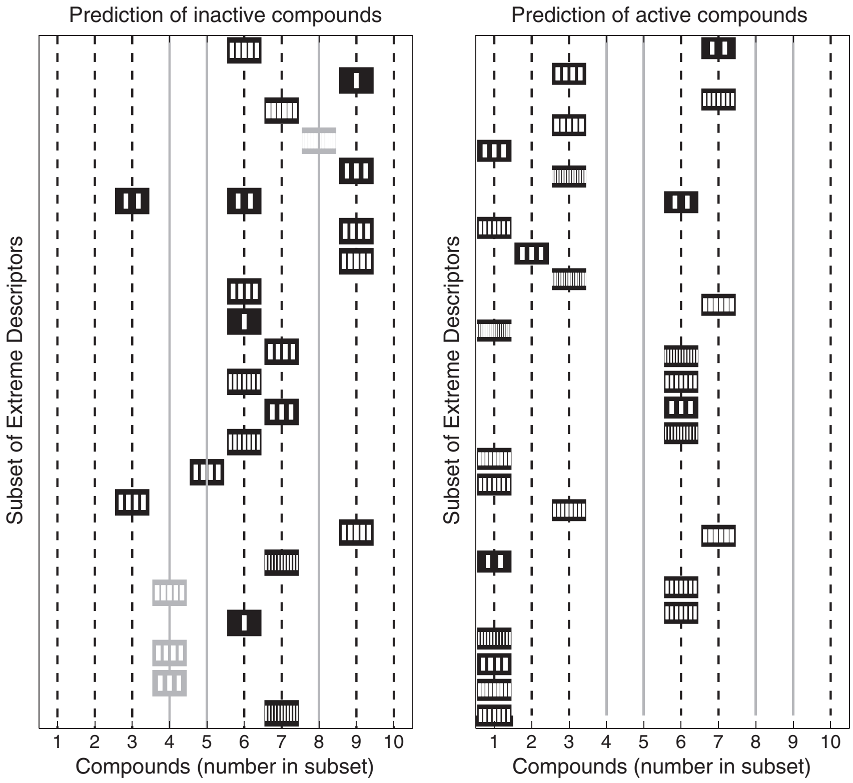

Figure 2. Predictions for chemical compounds from the skin sensitization data (134 sensitizers and 128 non-sensitizers) based on the distance-weighted discrimination classification model and descriptors with low variance. Compounds that are predicted as active and inactive by the distance-weighted discrimination classifier are denoted as solid gray and dashed black vertical lines. Predictions using low variance descriptors are denoted by small horizontal bars. White dots show the number of compounds in the training set that support the prediction. Compounds 1, 3, 6, and 7 from the right panel (original label is active) are strong candidates for being mislabeled.

accuracy for inactive compounds in both data sets, while there is still room for improvement for active compounds.

Additionally, we investigated the independent performance of both ED and DWD classifiers on the same set of compounds (those that contain extreme descriptors). It is possible to see that on the M_2 muscarinic receptor data set, the ED classifier performs significantly better compared with DWD. The ED classifier has the same overall accuracy as DWD on the skin sensitization data. While performing better on inactive compounds, the ED classifier fails to classify active compounds correctly. Note that in the next section, out of the 14 compounds that were misclassified by the ED predictor, five were identified as suspects (with one verified) of being mislabeled.

4.2. Identification of potentially mislabeled compounds

In this section, for each data set, we create a table that contains the list of compounds suspected to be mislabeled, their original label with corresponding confidence bound.

There are two main types of error that affect the label of each compound. The first corresponds to typographical or manual data entry errors. This type of error can be eliminated by confirming the label reported in the description of the original or repeated experiments. The second type comes from errors in the experiment and can be removed from the data set by repeating the experiment. Data entry errors are the easiest to confirm, but it is possible that the label reported in the original experiment is not correct.

A full list of mislabeled suspects, based on extreme descriptors without mislabeling, has six compounds from the skin sensitization data set and three compounds from the M_2 data. Out of the eight suspects, two compounds were verified to be mislabeled (one out of the six compounds from the skin sensitization data set and one out of the three compounds from the M_2 data set) because of data entry errors. Extreme descriptors with mislabeling produce similar list for skin sensitization data set and five additional suspects from M_2 data set.

Figure 2 summarizes the results over 20 randomly selected (10 inactive, left panel, and 10 active, right panel) leave-one-out folds. The detailed explanation of the format and interpretation of this figure are identical to Figure A1 and provided in Label prediction using extreme descriptors section of the Appendix. Some of the compounds on the left panel are predicted as active by the DWD classifier (solid gray horizontal lines). At the same time, it is possible to see that almost all predictions by extreme descriptors correspond with the predictions by the DWD. Out of the three compounds that predicted as active, only compound 8 has large number of compounds that support the prediction; however, all these compounds contain the same extreme descriptor. The situation is different for active compounds (right panel). In this subset, there are six compounds that are predicted as inactive by the DWD classifier. Only compounds 1, 3, 6, and 7 have a reasonable number of compounds from the training set (white dots) that support the inactive label prediction. The list of all six mislabeled suspects is presented in Table II. All suspects were checked for data entry errors, and compound

Table II. Mislabeled suspects (sorted by label confidence) from the skin sensitization and M_2 muscarinic receptors data sets

Data	Compound number	Original label	Label confidence	Verified mislabeling
Skin sens.	246	Active	0.06	
	43	Active	0.06	
	85	Active	0.07	**Yes**
	212	Inactive	0.08	
	82	Active	0.08	
	258	Active	0.10	
M_2	426	Active	0.06	**Yes**
	317	Active	0.11	
	347	Active	0.11	

Label confidence is based on the upper bound of 90% confidence interval for misclassification probability. Two of the suspects have been independently verified as mislabeled. sens., sensitization.

85 (compound 3 in the right panel of Figure 2) was verified to be mislabeled.

The M_2 muscarinic receptor data set that we considered contains 444 compounds (179 actives and 265 inactives) and 1116 descriptors. Table II contains the list of three active compounds that are predicted as inactive with a significant amount of support. Compound 426 was confirmed to be mislabeled because of a data entry error.

An interesting phenomenon happens when we allow mislabeling during the selection process of extreme descriptors. There are a group of six active compounds (9, 26, 333, 395, 409, and 426) that are predicted as inactive by extreme descriptors, and each of them has five active compounds that disqualify the prediction. Further investigation showed that compounds from this group are very similar to each other in the subspace generated by selected extreme descriptors. Additionally, active compounds that disqualify the prediction for each of the six compounds are remaining compounds from the group. In other words, they disqualify each other. The next logical step was to put the whole group into the test set and predict all of them at the same time. When all six compounds were predicted at the same time, only active compound 64 disqualified the prediction. Also, confidence of the original label decreased approximately from 12% to 5%. This fact makes them good candidates for being mislabeled, especially because compound 426 was verified to be mislabeled.

5. CONCLUSIONS

In this paper, we introduced the term “extreme predictors,” that is, descriptors that have a constant value for the majority of the compounds of the data set and only for a small fraction of compounds the values of these descriptors could vary. In the more specific version of extreme descriptors, the compounds from this small fraction must belong to the same activity class. Although extreme descriptors have very low total variance and are often discarded before modeling, we showed that they alone contain significant amounts of information and can be used for improving the quality of prediction. We also showed on two case studies that extreme descriptors can be successfully used for the identification of possibly mislabeled compounds.

Acknowledgements

The authors gratefully acknowledges the financial support from NIH (grant GM66940), NSF (grant 1016441), and EPA (RD 83382501 and R832720). The authors declare that they have no conflict of interest.

REFERENCES

1. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 2010; **50**: 1189–204.
2. Young D, Martin T, Venkatapathy R, Harten P. Are the chemical structures in your QSAR correct? *QSAR Comb. Sci.* 2008; **27** (11-12): 1337–1345.
3. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Oprea TI. WOMBAT: world of molecular bioactivity. In *Cheminformatics in Drug Discovery*, Oprea TI (ed). Wiley-VCH: New York, 2004; 223–239.
4. Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N, Hadaruga D, Moldovan R, Fulias A, Mracec M, Oprea TI. WOMBAT and WOMBAT-PK: bioactive databases for lead and drug discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, Schreiber SL, Kapoor TM, Wess G (eds). Wiley-VCH: New York, 2007; 760–786.
5. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV. Combinatorial QSAR modeling of chemical toxicants tested against tetrahymena pyriformis. *J. Chem. Inf. Model.* 2008; **48**: 766–784.
6. Wold S, Eriksson L, Clementi S. *Chemometric Methods in Molecular Design*. Wiley-VCH Verlag GmbH: Weinheim, Germany, 2008, 309–338.
7. Jolliffe IT. *Principal Component Analysis* (2nd edn). Springer: New York, 2005.
8. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* 2001; **58**(2): 109–130.
9. Tibshirani R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Series B* 1994; **58**: 267–288.
10. Bi J, Bennett K, Embrechts M. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.* 2003; **3**: 1229–1243.
11. Meier L, Van De Geer S, Bühlmann P. The group LASSO for logistic regression. *J. R. Stat. Soc. Series B (Statistical Methodology)* 2008; **70**: 53–71.
12. Brodley CE, Friedl MA. Identifying mislabeled training data. *J. Artif. Intell. Res.* 1999; **11**: 131–167.

13. Brodley CE, Friedl MA. Identifying and eliminating mislabeled training instances. In *AAAI/IAAI*. AAAI Press: Palo Alto, California, USA, 1996; 799–805.
14. Zhang W, Rekaya R, Bertrand K. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. *Bioinformatics (Oxford, England)* 2006; **22**: 317–25.
15. Joseph S, Robbins K, Zhang W, Rekaya R. Effects of misdiagnosis in input data on the identification of differential expression genes in incipient Alzheimer patients. *Silico Biol.* 2008; **8**: 545–54.
16. Gamberger D, Lavrač N, Džeroski S. Noise elimination in inductive concept learning: a case study in medical diagnosis. In *Algorithmic Learning Theory*, Arikawa S, Sharma A (eds), Lecture Notes in Computer Science, vol. 1160. Springer Berlin/Heidelberg, 1996; 199–212.
17. Wilson DR, Martinez TR. Instance pruning techniques. In *Machine Learning: Proceedings of the Fourteenth International Conference*. Morgan Kaufmann: San Francisco, CA, USA, 1997; 404–411.
18. Zeng X, Martinez T. An algorithm for correcting mislabeled data. *Intell. Data. Anal.* 2001; **5**: 491–502.
19. Teng C-M. Correcting noisy data. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; 239–248.
20. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat. Chem. Biol.* 2015; **11**(8): 535–535.
21. Benigni R, Giuliani A. Computer-assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Env. Heal. A* 1988; **25**: 135–48.
22. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller K-R. Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* 2009; **49**: 2077–81.
23. Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol. Appl. Pharm.* 2015; **284**(2): 262–272.
24. Quinlan JR. Induction of decision trees. *Mach. Learn.* 1986; **1**: 81–106.
25. Reduced murine local lymph node assay: an alternative test method using fewer animals to assess the allergic contact dermatitis potential of chemicals and products, Interagency Coordinating Committee on the Validation of Alternative Methods, Research Triangle Park, 2009.
26. Kuz'min VE, Artemenko AG, Muratov EN, Polischuk PG, Ognichenko LN, Liahovsky AV, Hromov AI, Varlamova EV. Virtual screening and molecular design based on hierarchical QSAR technology. In *Recent Advances in QSAR Studies*, Puzyn T, Leszczynski J, Cronin MT (eds), Challenges and Advances in Computational Chemistry and Physics, vol. 8. Springer: Netherlands, 2010; 127–176.
27. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, Germany, 2000.
28. Casella G, Berger R. *Statistical Inference*. Duxbury Resource Center: Pacific Grove, CA, 2001.
29. Efron B. R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture). *Stat. Sci.* 1998; **13**: 95–122.
30. Schweder T, Hjort NL. Confidence and likelihood. Large structured models in applied sciences; challenges for statistics. *Scand. J. Stat.* 2002; **29**: 309–332.
31. Hannig J, Xie M-g. A note on Dempster–Shafer recombination of confidence distributions. *Electron. J. Stat.* 2012; **6**: 1943–1966.
32. Hannig J. Generalized fiducial inference via discretization. *Stat. Sinica* 2013; **23**: 489–514.
33. Matthews BW. Comparison of the predicted and observed secondary structure of {T4} phage lysozyme. *Biochim. Biophys. Acta. (BBA) - Protein Structure* 1975; **405**(2): 442–451.
34. Marron JS, Todd M, Ahn J. Distance Weighted Discrimination. *J. Am. Stat. Assoc.* 2007; **102**: 1267–1271.
35. Hannig J. On generalized fiducial inference. *Stat. Sinica* 2009; **19**: 491–544.
36. de Haan L, Ferreira A. *Extreme Value Theory: An Introduction* (1st edn). Springer: New York, 2006.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

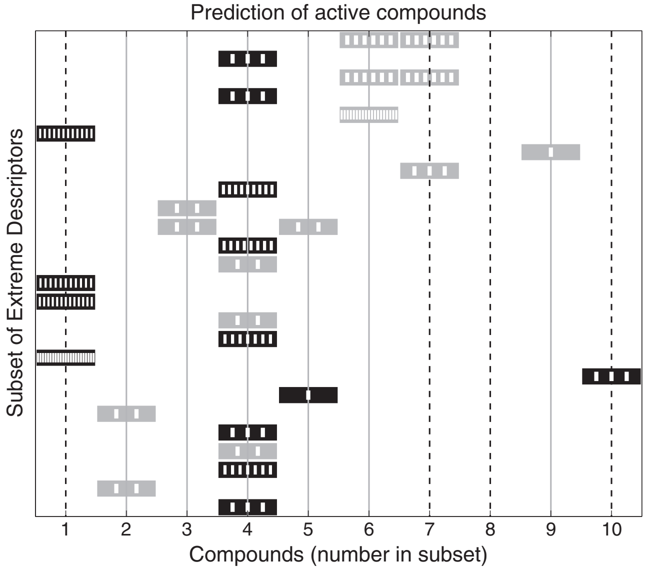

Figure A1. Example illustrates prediction of a subset of 10 compounds taken from the M_2 muscarinic receptor data set with DWD classification model and with extreme descriptors. Compounds that are predicted as active and inactive by DWD model are denoted as solid gray and dashed black vertical lines. Predictions by extreme descriptors are represented by small horizontal bars of appropriate color. White dots inside the horizontal bars show the number of compounds in the training set that support the prediction. Compound 1 with top two extreme descriptors is also shown in Figure 1(B) and 1(C). Compound 6 with third from the top extreme descriptor is shown in Figure 1(D).

APPENDIX A

A.1. Label prediction using extreme descriptors

The idea of how the predictions from the individual extreme descriptors can be combined is illustrated in Figure A1. It gives a simultaneous view of the classification results produced by two different methods. It shows the predictions of active compounds from the test set by the distance-weighted discrimination classifier (DWD) [34] and the set of extreme descriptors. To appropriately handle distributional irregularities that could significantly affect the DWD model, for example, skewness as shown in Figure 1(A), the MinSkew transformation (described in MinSkew transformation section) was applied to each descriptor in the data set. However, because the majority of extreme descriptors are binary, the MinSkew transformation will have almost no effect on their predictions. The data illustrated in Figure A1 are a small subset of 10 compounds taken from the M_2 muscarinic receptor data set (described in Section 4), chosen for illustrative purposes. In Figure A1, the horizontal axis represents compounds from the test set, and the vertical axis represents extreme descriptors identified in the training set. If DWD alone predicts a compound as active (inactive), it will be shown as a solid gray (dashed black) vertical line. Ideally, all compounds in Figure A1 should be predicted as active and represented by solid gray vertical lines. For example, compound number 6 is predicted as active (correct with respect to the given labeling), hence colored gray. On the other hand, compound 1 is predicted as inactive (incorrect prediction) and colored black. The second set of predictions is based only on the set of extreme descriptors. Those predictions are represented by short horizontal bars of appropriate color. The number of compounds from the training set that support the prediction is indicated by the number of white dots. Consider, for example, compound 6 (also shown as the black-triangle test case in Figure 1(D)). It has three extreme descriptors (short gray bars) that assign the active label and a total of 31 white dots (active compounds that support the active label assignment). Therefore, there are no reasons to dispute the labeling from the DWD model. Similarly, compound 1 (black-square test case in Figure 1(A) and (B)) is predicted as inactive by the DWD (black vertical line) and four extreme descriptors with a total of 59 supporting compounds. This compound is a good candidate to be checked for mislabeling or to be retested. Alternatively, compound 7 (black-diamond test compound in panel C of Figure 1) is predicted as inactive by the DWD classifier but assigned to the originally reported class by the set of extreme descriptors. In this case, there is a disagreement between two predictions. There are also compounds that cannot be predicted by extreme descriptors. For example, compound 8 does not have extreme descriptors that contain useful information. At the same time, compound 5 has two extreme descriptors that make opposite predictions supported. These are seen to be spurious because there are only two and one supporting compounds from opposite classes. In this case, the prediction by extreme descriptors does not provide any useful information.

A.2. Derivation of the original label confidence

Let $X_1, \ldots, X_n$ be a sample from a Bernoulli distribution with parameter θ. Then $X = \sum X_i$ is a binomial random variable with parameters n and θ and with distribution function $F(x|\theta)$. We can construct a confidence interval $(0, u)$ for the parameter θ, where u is defined by

$$F(x|u) = \alpha \tag{a1}$$

Using the fact that

$$\sum_{i=0}^{k} \binom{n}{i} u^i (1-u)^{n-i} = (n-k)\binom{n}{k} \int\limits_{0}^{1-u} t^{n-k-1}(1-t)^k dt$$

Equation (a1) can be written as

$$\alpha = \frac{1}{B(n-k, k+1)} \int\limits_{0}^{1-u} t^{n-k-1}(1-t)^k dt \tag{a2}$$

where k is the observed value of X and $B(\cdot,\cdot)$ is the beta function. Relationship (a2) can be simplified for particular values of k. For example,

$$\alpha = \begin{cases} (1-u)^n, & \text{if } k = 0, \\ 1 + (1-u)^{n-1}(u - 1 - nu), & \text{if } k = 1, \\ \frac{2(u-1)^2 - (1-u)^n(2-4u+2nu+2u^2-3nu^2+n^2u^2)}{2(u-1)^2} & \text{if } k = 2 \end{cases} \tag{a3}$$

Solving Equation (a3) for the unknown u produces the upper bound of the $1-\alpha$ confidence interval for the parameter θ.

This idea can be directly applied to the situation where we use extreme descriptors to predict the label of a chemical compound. For example, to construct a 95% confidence interval (0, 0.1), we would have to observe $n = 29$ compounds that support an active label and $k = 0$ compounds that support an inactive label. This also means that if we observe $n = 29$ and $k = 0$, then the original label confidence is bounded by 10%.

The estimate of the threshold n or the upper bound u of the confidence interval might be too conservative. We can use ideas of generalized fiducial inference [32,35] to find a $1-\alpha$ confidence interval for θ based on the half-corrected confidence distribution [29–31], which has good small and large sample properties [35]. For observed k and n, the half-corrected confidence distribution is given by

$$\frac{H(\theta,k) + H(\theta,k-1)}{2} \tag{a4}$$

where $H(\theta,k) = P(X > k|\theta)$ and X has a binomial (n,θ) distribution. Note that the distribution (a4) is the 50–50 mixture of the beta(k, n–k+1) and beta(k+1, n–k) distributions. In the case where $k = 0$, distribution (a4) becomes a mixture with components that take a value zero and beta(1, n) with equal probabilities. The upper bound of the confidence interval for $k = 0$ can be obtained from the following equation:

$$1 - \alpha = \frac{1}{2} + \frac{1}{2}\int_0^u (1-t)^{n-1} dt$$

which yields

$$u = 1 - (2\alpha)^{1/n} \tag{a5}$$

This equation can be also solved for n to obtain the number of compounds n required to achieve upper bound u with confidence $1-\alpha$. For example, to achieve a 10% upper bound on the probability that extreme descriptors made incorrect prediction with 95% confidence, it is required to observe $n = 22$.

The upper bound obtained using (1) is less conservative than the bound obtained from (a3) and will be used in the analysis of real chemical data.

A.3. Selection of extreme descriptors

This section describes the algorithm of the selection process of the extreme descriptors. The algorithm is designed for the general version of extreme descriptors that allows for a certain level of mislabeling in the data. For illustrative purposes, in this paper, we consider 8% of mislabeling [3]. Note that descriptors with zero variance in one of the classes are the special case of extreme descriptors with no mislabeling.

The following algorithm produces the list of extreme descriptors. A descriptor that satisfies all conditions is called an extreme descriptor. The conditions are structured as a sequential filter, where the input for each step is the output of the previous step.

(1) The median should be the same for both classes.
(2) All compounds should have the value of the descriptor on one side of the median, that is, greater or equal (smaller or equal) than the median.
(3) Proportion of possibly mislabeled compounds p_{Miss} should be smaller than 8%. The calculation of p_{Miss} is carried out in the following way:
    - For each class, calculate the number of compounds that are different from the median $\left(N_{act}^{diff} \text{ and } N_{inact}^{diff}\right)$.
    - Calculate the proportion p_{Miss} of possibly mislabeled compounds

$$p_{Miss} = \frac{\min\left(N_{act}^{diff}, N_{inact}^{diff}\right)}{N_{act}^{diff} + N_{inact}^{diff}}$$

(4) Values of possibly mislabeled compounds should be within the range of the other class. For example, if $N_{act}^{diff} > N_{inact}^{diff}$, then values of inactive compounds that are different from the median should be within minimum and maximum of the values of active compounds that are different from the median.

(5) Prediction is made only if the value of the new compound is also within the same range as in step 4.

A.4. MinSkew transformation

Because most statistical procedures, including DWD, are sensitive to gross changes in scale and skewness of marginal distributions, predictor transformation is important. We propose a transformation called *MinSkewness* to make the descriptors more amenable to statistical analysis, that is, closer to Gaussian.

The algorithm for the proposed transformation follows:

(1) For every descriptor x_i, decide if it is binary.

(a) For every non-binary descriptor $x_i \in \mathbb{R}^n$, calculate the sample skewness

$$g(x_i) = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^3}{\left(\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2\right)^{3/2}}$$

Then apply the family of shifted log transformations

$$x'_{ij} = \begin{cases} \log\left(x_{ij} - \min(x_i) + \alpha r_i\right) & g(x_i) > 0, \\ \log\left(\max(x_i) - x_{ij} + \alpha r_i\right) & g(x_i) < 0, \\ x_{ij} & g(x_i) = 0 \end{cases}$$

where $r_i = \max(x_i) - \min(x_i)$ and $i = 1, \ldots, n$. Choose the parameter α to minimize the absolute sample skewness $g(x')$.

(b) If the descriptor x_i is binary, no transformations are applied because it will not change the skewness, that is, $x'_i = x_i$.

(2) Standardize every descriptor x' in the following way:

$$x''_{ij} = \frac{x'_{ij} - \bar{x}'_i}{s_i}$$

where $\bar{x}'_i$ and s_i are the sample mean and standard deviation of the transformed descriptor x'_i, respectively.

(3) If $|x''_{ij}| > L$, where L is the threshold defined in the succeeding texts, truncate descriptors $x'''_{ij} = sign\left(x''_{ij}\right) L$.

There are many descriptors that have very small sample standard deviation because of the majority of the compounds sharing the same value for that descriptor. Standardization of such descriptors by the sample standard deviation as part of the transformation proposed in this section will significantly magnify its values. One possible remedy is truncation of the descriptor values.

Let $\mathbf{x} = \left(x_1^T, \ldots, x_m^T\right)^T$ be the column vector of stacked descriptor values, where $x_i = (x_{i1}, \ldots, x_{in})^T$ is the ith standardized descriptor with zero mean. If the absolute value of the element of $\mathbf{x}$ is greater than a certain threshold, that is, $|x_{ij}| > L$, then $x'_{ij} = sign(x_{ij})L$.

A reasonable value of the threshold L is based upon the following fact. Let $X_1, X_2, \ldots$ be independent, identically distributed standard normal random variables. Define the maximum of k random variables by $M_k = \max(X_1, \ldots, X_k)$. By a classical result from the extreme value theory [36], there exist real constants $a_k > 0$ and b_k such that

$$\frac{M_k - b_k}{a_k} \xrightarrow{D} Y \qquad \text{as } k \to \infty$$

where Y has the standard Gumbel distribution and

$$b_k = (2\log k)^{1/2} - \frac{\log\log k + \log(4\pi)}{(2\log k)^{1/2}},$$
$$a_k = (2\log k)^{-1/2}$$

A reasonable large value is thus based on the 95th percentile of the standard Gumbel distribution p_{95}. Take the corresponding threshold to be

$$L = p_{95} a_k + b_k$$

where $k = mn$.