J Comput Aided Mol Des (2014) 28:941–950
DOI 10.1007/s10822-014-9778-3

ChemStable: a web server for rule-embedded naïve Bayesian learning approach to predict compound stability

Zhihong Liu · Minghao Zheng · Xin Yan · Qiong Gu · Johann Gasteiger · Johan Tijhuis · Peter Maas · Jiabo Li · Jun Xu

Received: 8 March 2014 / Accepted: 9 July 2014 / Published online: 17 July 2014

Abstract Predicting compound chemical stability is important because unstable compounds can lead to either false positive or to false negative conclusions in bioassays. Experimental data (COMDECOM) measured from DMSO/H_2O solutions stored at 50 °C for 105 days were used to predicted stability by applying rule-embedded naïve Bayesian learning, based upon atom center fragment (ACF) features. To build the naïve Bayesian classifier, we derived ACF features from 9,746 compounds in the COMDECOM dataset. By recursively applying naïve Bayesian learning from the data set, each ACF is assigned with an expected stable probability (p_s) and an unstable probability (p_{uns}). 13,340 ACFs, together with their p_s and p_{uns} data, were stored in a knowledge base for use by the Bayesian classifier. For a given compound, its ACFs were derived from its structure connection table with the same protocol used to drive ACFs from the training data. Then, the Bayesian classifier assigned p_s and p_{uns} values to the compound ACFs by a structural pattern recognition algorithm, which was implemented in-house. Compound instability is calculated, with Bayes' theorem, based upon the p_s and p_{uns} values of the compound ACFs. We were able to achieve performance with an AUC value of 84 % and a tenfold cross validation accuracy of 76.5 %. To reduce false negatives, a rule-based approach has been embedded in the classifier. The rule-based module allows the program to improve its predictivity by expanding its compound instability knowledge base, thus further reducing the possibility of false negatives. To our knowledge, this is the first in silico prediction service for the prediction of the stabilities of organic compounds.

Keywords Naïve Bayesian classifier · Stability prediction · Atom center fragment · Knowledge rule

Zhihong Liu and Minghao Zheng have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9778-3) contains supplementary material, which is available to authorized users.

Z. Liu · M. Zheng · X. Yan · Q. Gu · J. Xu (✉)
Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-sen University, 132 East Circle at University City, Guangzhou 510006, China
e-mail: junxu@biochemomes.com

J. Gasteiger
Molecular Networks GmbH, Henkestraße 91, 91052 Erlangen, Germany

J. Tijhuis · P. Maas
Specs, Kluyverweg 6 (Building 65), 2629 HT Delft, The Netherlands

J. Li
SciNet Technologies, 9943 Fieldthorn St., San Diego, CA 92127, USA

Introduction

In pharmaceuticals industry, storing compounds for a long time raises concerns about chemical stability. Unstable compounds can lead to either false positive or to false negative conclusions in bioassays, and thus leading to incorrect inferences concerning structure–activity relationships [1]. An early stability screening strategy was introduced to identify labile chemotypes [2, 3]. However, this is costly. A number of methods have been reported for estimating compound chemical stability [4–6]. Stability is affected by a number of factors: chemical features, solvent types, temperature, humidity, freeze–thaw resistance,

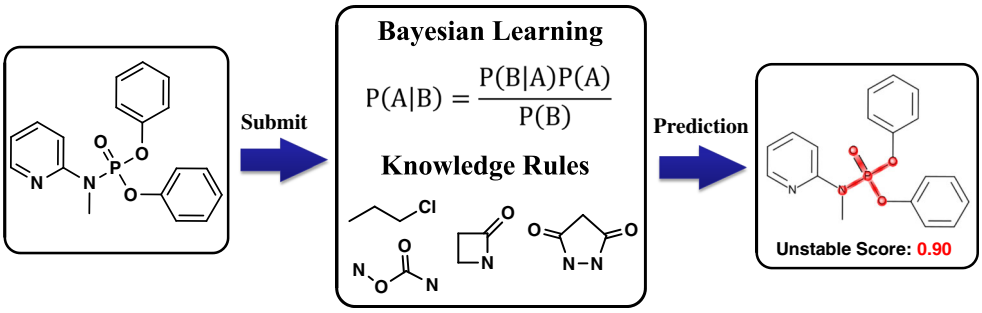

Fig. 1 The flowchart of the ChemStable for chemical stability prediction

storage time, etc. [3, 7–10]. Recently, Popa-Burke et al. [11] reported that initial purity was the only factor that had a clear effect on stability. It is better to estimate stability by deriving models from experimental data. The COMDECOM project was established for this purpose [12]. In previous work, a predictive stability model was built, based upon the COMDECOM data measured from DMSO/H_2O solutions stored at 50 °C for 105 days; the model was able to predict stability with 70 % accuracy. It is a general phenomenon in whole molecule based modeling that the greater the chemical diversity of the investigated compounds, the smaller the chance that SAR models exist and can be uncovered. On the other hand, the information content of an SAR model will increase as the boundaries of the chemical space and the diversity of the compounds under investigation increases [13]. The stability of a compound can be determined by a localized structural fragment. Therefore, modeling with fragments might improve the success rate [14]. To improve the success rate, we re-visited the data by applying a rule-embedded naïve Bayesian learning [15], and employed an atom center fragment (ACF) approach [16–18] to extract features from the experimental data for the learning process. To make full use of the COMDECOM data, we also derived rules, from the ACF features, that reflect stability-structure relations and embed the rules into the naïve Bayesian classifier. A web server was developed based on the embedded Bayesian classifier. The structures can be submitted to our web server by drawing a molecule online or by uploading a library in SDF file format. The prediction will be run in the backend and the unstability score will be provided and the predicted potential reactive site will be highlighted. The flowchart of the web server is illustrated in Fig. 1.

Materials and methods

Data set

The COMDECOM data contain 12,810 structurally diverse compounds and their stability data (which were measured with unified protocols [12]). The compounds were tested in

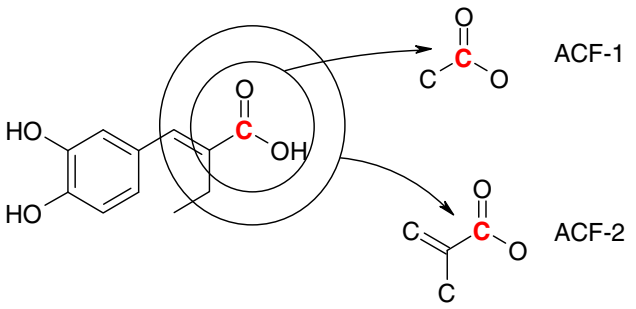

Fig. 2 Examples for ACF-1 and ACF-2: the atoms in red are center atoms

DMSO and DMSO/H2O (20 % v/v of water) profiles. The data used in this work were measured in DMSO/H2O profiles. The compound purities and the stability indices were measured at 50 °C for 0, 14, 35, and 105 days. Here, we did a careful analysis of the problems in the dataset. The records containing erroneous purity profiles, such as a compound concentration c_t at time t that is higher than c_{t-1} by +5 %, are removed. Duplicated records or those concerning salts (or non-covalent complexes) were discarded. After these filtering processes, 10,160 compound records remained. A compound purity measured at the 0th day is denoted as p_0, which usually ranges from 90 to 100 %, the purity at the 105th day is denoted as p_{105}, which is expected to be less than p_0 due to possible degradation. The instability of a compound can be quantitatively represented as the ratio of the two purities (i.e., p_{105}/p_0). 6,442 compounds are considered stable because $p_{105}/p_0 \geq 80$ %; 3,304 compounds are considered unstable because $p_{105}/p_0 \leq 70$ %. Compounds having a p_{105}/p_0 between 70 and 80 % were taken out of consideration. To summarize, 9,746 compounds were used to build the predictive stability model.

Method for generating features

An ACF is a substructure, which consists of a center atom and environmental layers (1, 2, 3, etc.); hydrogen atoms are excluded. Therefore, ACF-1 contains a center atom and the atoms that are covalently bonded to the center atom; ACF-

2 contains an ACF-1 and the atoms that covalently bonded to the atoms in ACF-1; and so forth (Fig. 2).

Usually, an ACF-(n + 1) is larger than an ACF-n. Larger ACFs may result in more accurate predictions, but the predictivity could be limited by more restrictive pattern recognition requirements. All ACF-1–4 fragments have been generated. The optimized ACF level was determined experimentally, and is discussed in the "Results and discussion" section.

Method to calculate p_s and p_{ins}

ACF-count(i) is the number of ACF(i) appearances in different molecules. ACF(i) may appear in the same molecule several times, but ACF-count(i) only increases once. Let $s(i)$ be the ACF-count(i) in 6,442 stable molecules, and $uns(i)$ be the ACF-count(i) in 3,304 unstable molecules. For a given ACF(i), its expected probabilities of chemical stability and instability, denoted as $p_s(i)$ and $p_{uns}(i)$, can be calculated as follows:

$$p_{uns}(i) = \frac{uns(i) + 0.5}{uns + 1.0} \tag{1}$$

$$p_s(i) = \frac{s(i) + 0.5}{s + 1.0} \tag{2}$$

where uns represents the number of unstable compounds; and s represents the number of stable compounds. In our case, these figures are 3,304 and 6,442, respectively.

Bayesian classifiers

Bayesian learning is a probabilistic classification approach [19–24] based on Bayes' theorem:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \tag{3}$$

where P(A) is the initial degree of belief in A; P(B) is the initial degree of belief in B; P(A|B) is the degree of belief having accounted for B; and P(B|A) is the degree of belief having accounted for A. Therefore, the probability of compound instability (P) can be calculated as the following:

$$\begin{aligned}\mathrm{P} &= \mathrm{P}(+|\mathrm{p_{uns}}(1), \mathrm{p_{uns}}(2), \ldots, \mathrm{p_{uns}}(\mathrm{n})) \\ &= \frac{\mathrm{P}(\mathrm{p_{uns}}(1), \mathrm{p_{uns}}(2), \ldots, \mathrm{p_{uns}}(\mathrm{n})|+)\mathrm{P}(+)}{\mathrm{P}(\mathrm{p_{uns}}(1), \mathrm{p_{uns}}(2), \ldots, \mathrm{p_{uns}}(\mathrm{n}))}\end{aligned} \tag{4}$$

where $\mathrm{P}(p_{uns}(1),p_{uns}(2),\ldots,p_{uns}(n)|+)$ is the probability that a compound has a set of ACFs that are responsible for chemical instability. P(+) is the prior probability [see Eq. (11)], a probability derived from the training set, and $\mathrm{P}(p_{uns}(1), p_{uns}(2), \ldots, p_{uns}(n))$ is the marginal probability that given ACFs appear in the data set.

In a naïve Bayesian classifier, descriptors are assumed to be independent of each other. Thus, the probability of compound instability $\mathrm{P}(p_{uns}(1), p_{uns}(2),\ldots,p_{uns}(n)|-)$, and stability $\mathrm{P}(p_s(1),p_s(2),\ldots,p_s(n)|-)$, can be calculated as follows:

$$\begin{aligned}&\mathrm{P}(p_{uns}(1), p_{uns}(2), \ldots, p_{uns}(n)|+) \\ &= \mathrm{P}(p_{uns}(1)|+)\mathrm{P}(p_{uns}(2)|+)\ldots\mathrm{P}(p_{uns}(n)|+) \\ &= \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(p_{uns}(i)|+)\end{aligned} \tag{5}$$

$$\begin{aligned}&\mathrm{P}(p_s(1), p_s(2), \ldots, p_s(n)|-) \\ &= \mathrm{P}(p_s(1)|-)\mathrm{P}(p_s(2)|-)\ldots\mathrm{p}(p_s(n)|-) \\ &= \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(p_s(i)|-)\end{aligned} \tag{6}$$

where (5) and (6) can be estimated from the training set:

$$\mathrm{P}(p_{uns}(i)|+) = \frac{\mathrm{count}(\mathrm{p_{uns}}(\mathrm{i})|+)}{\mathrm{count}(+)} \tag{7}$$

$$\mathrm{P}(p_s(i)|-) = \frac{\mathrm{count}(\mathrm{p_s}(\mathrm{i})|-)}{\mathrm{count}(-)} \tag{8}$$

The experimental data were divided into unstable and stable classes. The stable probability Q can be calculated in (9):

$$\begin{aligned}\mathrm{Q} &= \mathrm{P}(-|\mathrm{p_s}(1), \mathrm{p_s}(2), \ldots, \mathrm{p_s}(\mathrm{n})) \\ &= \frac{\mathrm{p}(\mathrm{p_s}(1), \mathrm{p_s}(2), \ldots, \mathrm{p_s}(\mathrm{n})|-)\mathrm{P}(-)}{\mathrm{P}(\mathrm{p_s}(1), \mathrm{p_s}(2), \ldots, \mathrm{p_s}(\mathrm{n}))}\end{aligned} \tag{9}$$

Since P + Q = 1, we have:

$$\begin{aligned}\log\frac{\mathrm{P}}{\mathrm{Q}} &= \log\frac{\mathrm{P}}{1-\mathrm{P}} = \log\frac{\mathrm{P}(\mathrm{p_{uns}}(1),\mathrm{p_{uns}}(2),\ldots,\mathrm{p_{uns}}(\mathrm{n})|+)\mathrm{P}(+)}{\mathrm{P}(\mathrm{p_s}(1),\mathrm{p_s}(2),\ldots,\mathrm{p_s}(\mathrm{n})|-)\mathrm{P}(-)} \\ &= \log\frac{\prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P}(\mathrm{p_{uns}}(\mathrm{i})|+)}{\prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{P}(\mathrm{p_s}(\mathrm{i})|-)} + \log\frac{\mathrm{P}(+)}{\mathrm{P}(-)}\end{aligned} \tag{10}$$

$$\mathrm{P}(+) = \frac{uns}{s + uns} \tag{11}$$

$$\mathrm{P}(-) = \frac{s}{s + uns} \tag{12}$$

where s and uns are the total numbers of stable and unstable compounds in the training data, respectively. Zero counts (i.e., ACFs only appear in the stable class or vice versa) were treated by applying the following Laplacian correction [25–29]:

$$\mathrm{P}(\mathrm{p_{uns}}(\mathrm{i})|+) = \frac{\mathrm{count}(\mathrm{p_{uns}}(\mathrm{i})|+) + 0.5}{\mathrm{count}(+) + 1} \tag{13}$$

$$\mathrm{P}(\mathrm{p_s}(\mathrm{i})|-) = \frac{\mathrm{count}(\mathrm{p_s}(\mathrm{i})|-) + 0.5}{\mathrm{count}(-) + 1} \tag{14}$$

Table 1 Diversity analyses for the COMDECOM, Drug Bank, and WDI databases

Data Set	Compounds	Scaffolds	Diversity (scaffolds/compounds) (%)
COMDECOM	9,746	6,117	62.76
Drug bank	6,516	2,784	42.70
WDI	70,555	24,557	34.80

Missing values (i.e., the number of ACFs that appear in the query structure, but which cannot be found in the knowledge base) were skipped [25, 26, 28].

Notably, a compound is regarded as "unstable" if it has at least one unstable fragment. The corollary is that a compound is stable when every fragment in the molecule is a stable fragment. Therefore, an ACF fragment cannot be neglected due to its lower population, and all ACFs have to be taken into account for a naïve Bayesian learning.

Validating models

A tenfold cross validation scheme was employed to evaluate the accuracy and robustness of the Bayesian Classifiers. The performance of the classifier was assessed by the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Sensitivity (SE: the ability to correctly predict unstable compounds), specificity (SP: the ability to correctly predict stable compounds), and global accuracy (GA: the proportion of true predictions out of the entire population) can be calculated as follows to validate our predictive models: [20, 30]

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{16}$$

$$\mathrm{GA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{17}$$

The receiver operating characteristic (ROC) [21] curve was plotted. The area under the ROC curve (AUC) was calculated for validating the models.

Results and discussion

Diversity and scaffold analyses

A scaffold-based classification approach [31] (SCA) was employed to compare the structural diversities of the COMDECOM, Drug Bank [32], and WDI [33] databases. Table 1 indicates that the ratio of scaffolds and compounds in COMDECOM is higher than those of the Drug Bank and WDI databases, thus pointing to a higher diversity.

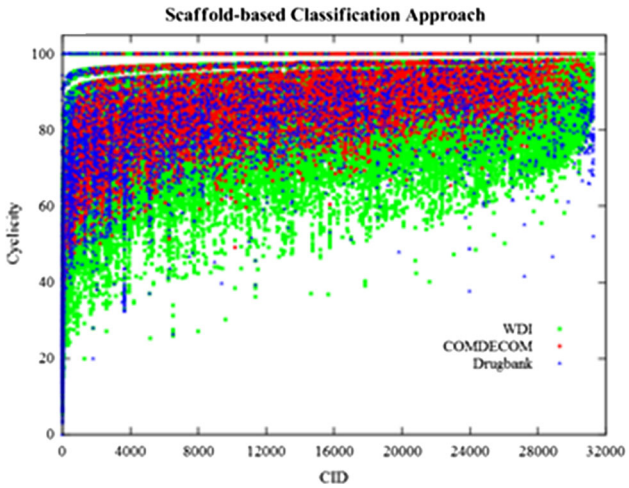

Fig. 3 The chemical scaffold distributions for the COMDECOM, WDI, and Drug Bank databases through scaffold-based classification approach (the number of compounds in COMDECOM library: 9,746; the number of compounds in Drugbank: 6,516; the number of compounds in WDI: 70,555)

The structural diversity distributions of the COMDECOM, WDI, and Drug Bank databases are shown in an SCA plot (Fig. 3). Cyclicity is the metric for the cyclic degree of a compound (a higher cyclicity value means that the compound has fewer side chains). Each compound has a CID (cluster ID) representing a scaffold. The CID value is associated with a compound's complexity. More complicated compounds have higher CID numbers. This analysis indicates that the chemical scaffolds of COMDECOM library cover the scaffolds of drug-like or lead-like compounds.

9,746 compounds have 6,117 scaffolds. 3,876 scaffolds are derived from stable compounds; 1,880 scaffolds are derived from unstable compounds. 361 scaffolds are derived from both stable and unstable compounds. The distributions of stable and unstable compounds in a scaffold family for larger scaffold families (i.e., family sizes >10) from the COMDECOM data are depicted in Fig. 4. Most of the larger compound families have both stable and unstable compounds. In Fig. 4, a few scaffold families, such as scaffolds 40–41 and 47–48, contain only unstable compounds; scaffolds 15–16, 25, 28, 37, and 46 contain only stable compounds. However, it is unclear whether these cases are due to unbalanced sampling or result from the scaffolds themselves being unstable or stable.

Finding the optimized ACF-level

Four levels of ACFs (ACF-1, ACF-2, ACF-3, and ACF-4) have been derived from the COMDECOM data, and trailed via the naïve Bayesian learning process to see which level

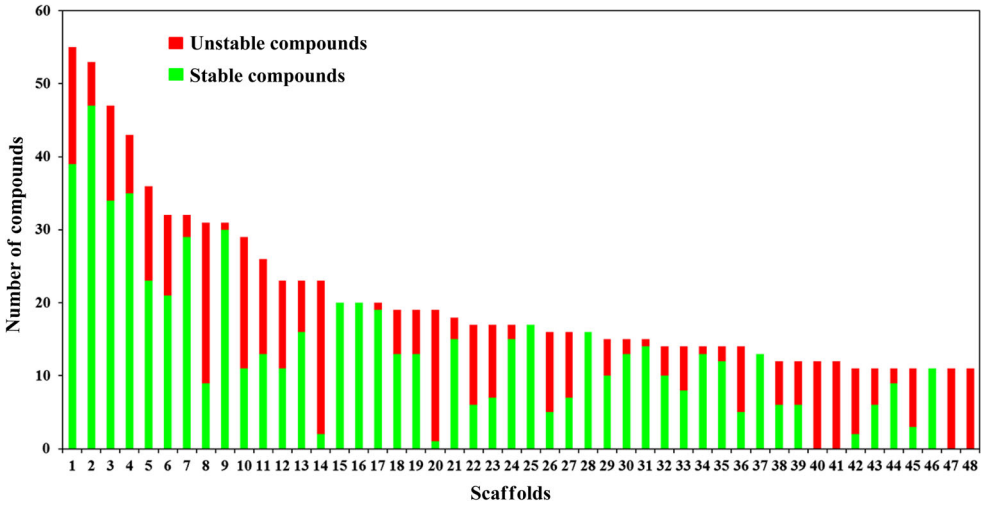

Fig. 4 The distributions of stable and unstable compounds in a scaffold family for larger scaffold families (family sizes >10 members) in COMDECOM data

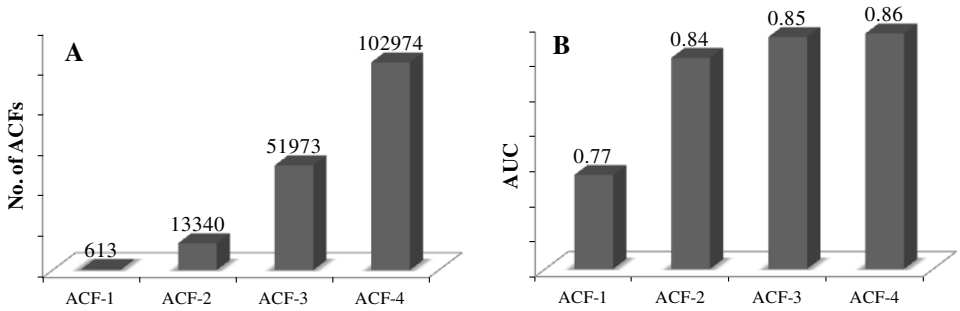

Fig. 5 The relation between Bayesian learning performance and the ACF-levels. **a** The relation of the number of ACFs with ACF-level (ACF-1, ACF-2, etc.). **b** The relation of the AUC of the Bayesian model tenfold cross validation with ACF-level

generates the best performance (in terms of predictivity). To figure out which level of ACF features are the best for Bayesian learning, we employed a tenfold cross validation scheme to validate the model's performance (measured in AUC), when using ACF features at different ACF-levels. The results are depicted in Fig. 5.

As shown in Fig. 5a, the number of ACFs increases exponentially along with the ACF-level. However, predictivity does not increase significantly along with the ACF-level (Fig. 5b). There is only one performance jump from ACF-1 to ACF-2. Consequently, ACF-2 features are the optimized ones for our naïve Bayesian models.

The tenfold cross validation also produced ROC curves (Fig. 6), which represent the ratio of true positive predictions against false positive predictions. As shown in Fig. 6, the model using ACF-1 features generates a poorer ROC curve. However, the model using ACF-2–4 features generate similar ROC curves, which are all superior to the ROC curve generated by ACF-1 features. Therefore, both ROC and AUC parameters indicate that the optimized ACF-level is 2.

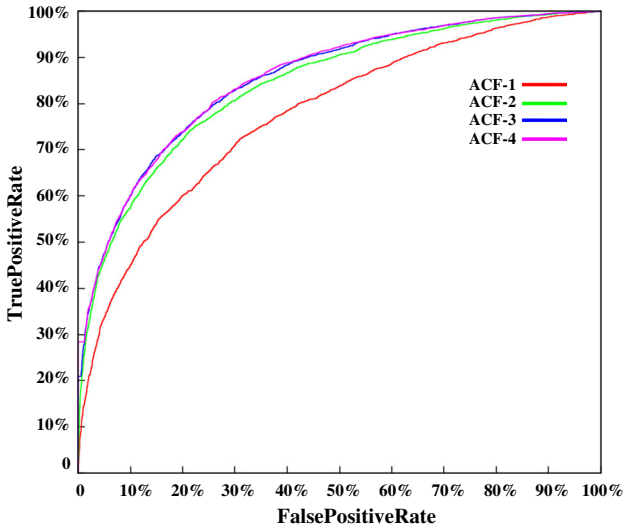

Fig. 6 The ROC curves for the tenfold cross validations of the Bayesian models with ACF features at ACF-levels 1–4

Robust validations

A Y-scrambling validation protocol was designed and described in supplementary material S1 to confirm that the classifier cannot be generated randomly [34]. The averages of SE, SP, AUC, and GA for the 100 scrambled models are 0.41, 0.61, 0.50, and 0.54, respectively. The averages of SE, SP, AUC, and GA for the 20 unscrambled models are 0.75, 0.77, 0.84, and 0.76, respectively (supplementary

Table 2 The profiles of the training set and testing set derived from COMDECOM

	Training set compounds	Test set compounds	Training set ACFs	Test set ACFs
Stable	4,572	2,272	8,424	5,957
Unstable	1,926	976	6,102	4,155
Ratio (stable/unstable)	2.37	2.33	1.38	1.43

material S2). Therefore, the validation parameters of unscrambled models are significantly more consistent and better than the ones with the scrambled models. This proves that our ACF-2 fragment based Bayesian model is not the result of chance correlation.

To further validate the performance of the classifier, additional tests were executed by dividing the COMDECOM dataset (containing 9,746 compounds) into two groups: the training set (about 2/3 of the compounds), and the testing set (about 1/3 of the compounds). The profiles of the two groups are listed in Table 2, which indicates that the ratio of stable to unstable compounds and the ratio of stable and unstable ACFs are almost equivalent. The values of SE, SP, AUC, and GA are 0.74, 0.77, 0.84, and 0.76, respectively. These results are consistent with the results of tenfold cross validations. Hence, the classifier is robust.

Rule-embedded Bayesian classification

It is impossible for an empirical model to attain 100 % success in predictive power. Noises, missed data, and incomplete data always exist. To make up for these deficiencies, we have collected 40 unstable fragments from references [4, 35]. We further analyzed the unstable ACF-2 features. By combining the unstable fragments collected from references and the unstable ACF-2 features, and by removing the duplicated fragments of the two sets, we concluded that 20 unstable fragments could be used to generate instability prediction rules (IPR) (supplementary material S3). Consequently, we built a rule-embedded naïve Bayesian classifier, whose flow chart is shown in Fig. 7. In order to compare our models to the previous published models [12], we have calculated the accuracy of tenfold cross validation and compared them with corresponding values from the previous publications in Table 3. An increase in 0.3 % might not large, however, it was the effect that 30 compounds that previously false negatives are now correctly predicted. As, on the other hand 30 is a small number compared to the 9,746 compounds of the dataset. It should be emphasized that the value of this work is that we developed the predictive models through a Bayesian approach based on the unique COMDECOM database and provided a web service. Rule-embedded

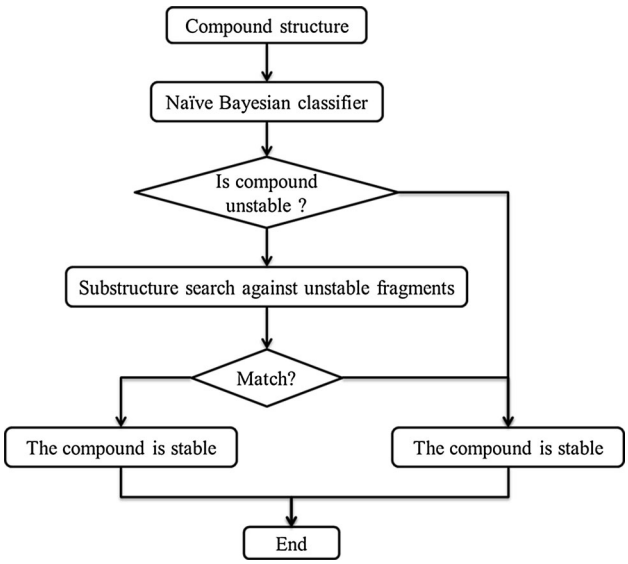

Fig. 7 The flow chart for the rule-embedded naïve Bayesian classifier

Table 3 Comparison of our models and previous models

Model method	Compounds	Tenfold CV (%)	Prediction service	Reactive site prediction
PLSR[a]	11,193	71.5	No	No
k-Nearest neighbors[a]	11,193	71.8	No	No
Decision trees[a]	11,193	70.1	No	No
Random forest[a]	11,193	72.9	No	No
Naïve Bayesian classifier[b]	9,746	76.2	Yes	Yes
Rule-embedded naïve Bayesian classifier[b]	9,746	76.5	Yes	Yes

[a] Previous published models [12]

[b] Our models in this work

further reduces the number of false negatives. To our knowledge, this is the first in silico service for the prediction of stabilities of organic compounds with the potential of highlighting the reaction site. The details of the web service can be seen in the next section.

Dependency of Bayesian score and the prediction accuracy

In this work, the bayesian score ranges from 0 to 1 and 0.5 is the cutoff value to distinguish the unstable or stable compounds. We analyzed the distribution of the Bayesian score and the number of true predictions (true positives and true negatives). As shown in Fig. 8, the majority of true positives (unstable compounds predicted as unstable) with a score >0.9

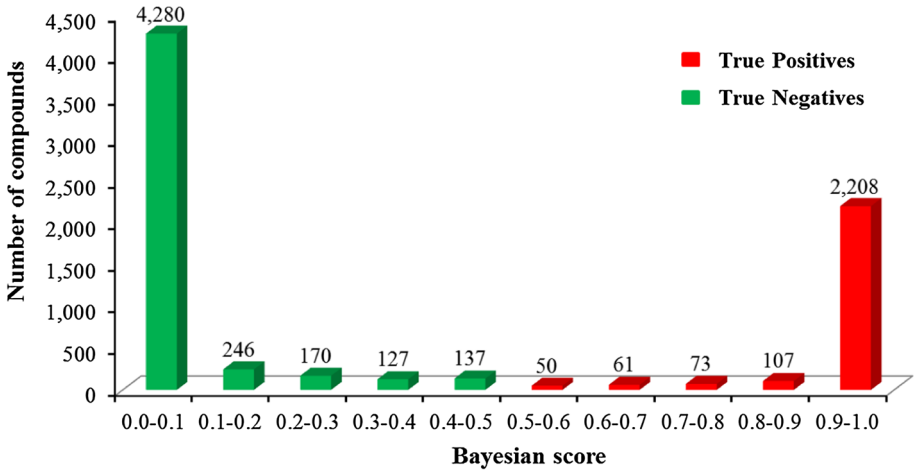

Fig. 8 The Bayesian score distributions of the true positives and true negatives

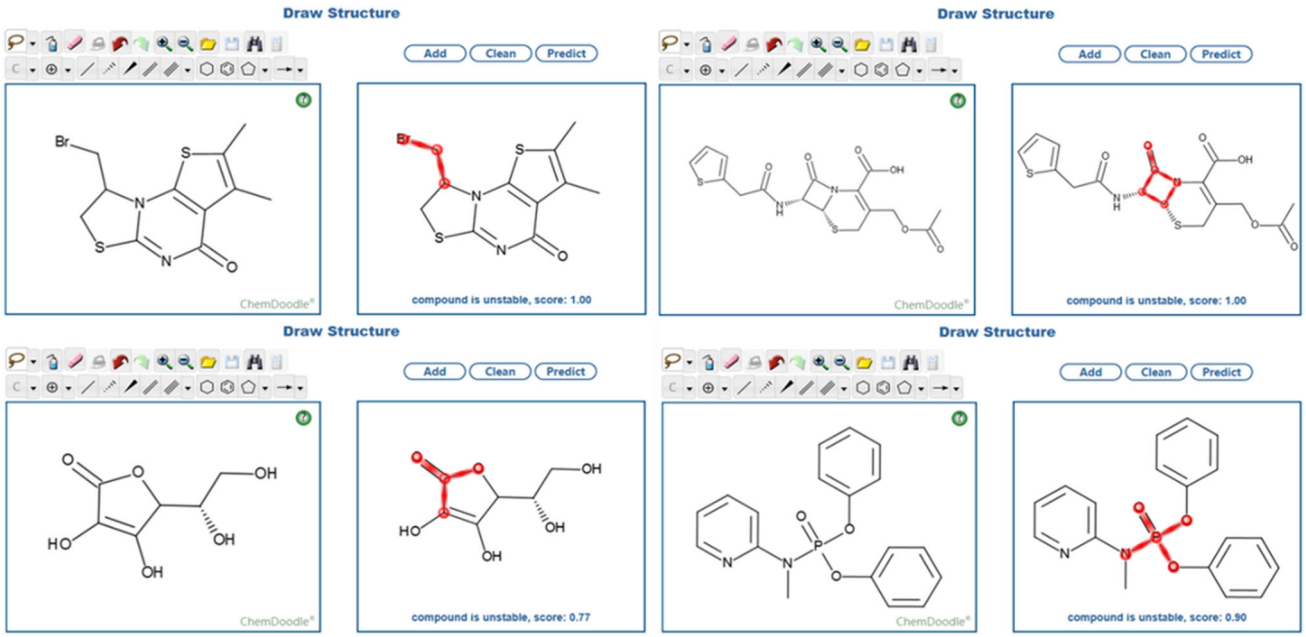

Fig. 9 Unstable features annotated by the ChemStable program. *Red* unstable features

and the majority of true negatives (stable compounds predicted as stable) give a score <0.1, which suggests the higher score corresponds to better accuracy for unstable compounds and the lower score correspond to better accuracy for stable compounds. For the compounds with a score in the "uncertainty area" between the classes (i.e. 0.4–0.6), the prediction is less reliable. Therefore, this is in agreement with the conclusion that the predictions close to the edge of the class have better accuracy than that in the "uncertainty area" [36].

Unstable fragment annotation and web-based server

Chemists may wonder which portion of the structure is responsible for instability when a compound is predicted by a program. We have developed a web-server, *ChemStable*, which employs the rule-embedded naïve Bayesian classifier to predict chemical stability and which highlights the corresponding unstable fragment's features. When a compound is predicted as unstable, the related unstable feature(s) can be annotated, as shown in Fig. 9.

ChemStable is implemented in HTML and Go language (an new language develop by Google), and accessible through a web browser. *ChemStable* adopts ChemDoodle (web.chemdoodle.com), an open source chemical structure drawing program, for a user to draw a chemical structure so as to predict the stability of an individual compound. The user can also upload a chemical compound library (SDF file format) to *ChemStable*, and predict the stabilities for all the compounds in the library. Figure 10 gives an example of the stability predictions for a chemical library.

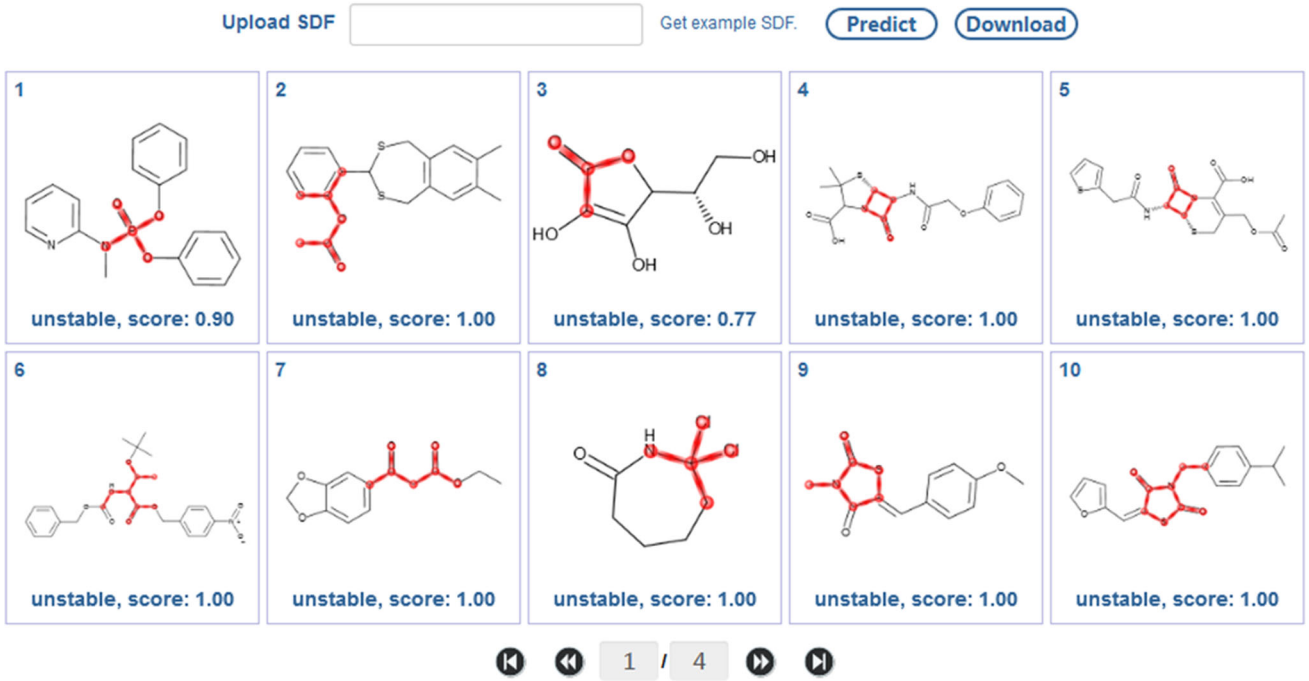

Fig. 10 Example of the stability predictions for a chemical library

Biodegradability and stability of COMDECOM dataset

For drug discovery, a compound stability is associated with not only chemical degradability, but also biodegradability. The biodegradability model has been reported by Vorberg [37], and is available online [38]. The biodegradability of the COMDECOM compounds was predicted by Vorberg's model. Sixty-seven COMDECOM compounds could not be predicted for the biodegradability in Vorberg's OCHEM program. Forty-six (0.72 %) COMDECOM chemically stable compounds were predicted as readily biodegradable and 6,365 (99.28 %) COMDECOM compounds were predicted as not readily biodegradable. Among the chemically unstable COMDECOM compounds, 40 (1.22 %) compounds were predicted as readily biodegradable and 3,228(98.78 %) compounds were predicted as not readily biodegradable (Table 4). The results indicate that most of COMDECOM compounds are not readily biodegradable and the unstable compounds are more likely to be readily biodegradable.

Table 4 Biodegradability of COMDECOM compounds predicted by Vorberg's approach

	Readily biodegradable	Not readily biodegradable
Stable dataset	46 (0.72 %)	6,365 (99.28 %)
Unstable dataset	40 (1.22 %)	3,228 (98.78 %)

Conclusion

To improve our chemical stability prediction model, we have re-visited the data of a previous publication and applied rule-embedded naïve Bayesian learning based upon ACF features, and achieved an AUC value of 84 % and a tenfold cross validation accuracy of 76.5 %. The classifier has been optimized and has experienced rigorous validations as well as robust tests. For compounds that were predicted as unstable, the program highlights the unstable structural features so that scientists can understand the mechanisms of instability. To reduce false negatives, a rule based approach was embedded in the classifier. The rule-based module allows the program to improve its predictivity by expanding its compound instability knowledge base. With an expandable knowledge base, we have the opportunity to further reduce false negatives. Also, the unstable feature highlighting function makes the predictions interpretable, and provides an opportunity for the user to examine whether false positives or negatives exist.

Acknowledgments This work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2012AA020307), the Guangdong Recruitment Program of Creative Research Groups, the National Natural Science Foundation of China (No. 81173470), and the Special Funding Program for the National Supercomputer Center in Guangzhou (2012Y2-00048/2013Y2-00045, 201200000037). The authors would like to thank Mr. Heming Xu of Columbia University

for his comments and corrections, which have improved the manuscript.

References

1. Di L, Kerns EH (2009) Stability challenges in drug discovery. Chem Biodivers 6(11):1875–1886. doi:10.1002/cbdv.200900061
2. Blaxill Z, Holland-Crimmin S, Lifely R (2009) Stability through the ages: the GSK experience. J Biomol Screen 14(5):547–556. doi:10.1002/cbdv.200900061
3. Cheng XH, Hochlowski J, Tang H, Hepp D, Beckner C, Kantor S, Schmitt R (2003) Studies on repository compound stability in DMSO under various conditions. J Biomol Screen 8(3):292–304. doi:10.1177/1087057103008003007
4. Waterman KC, Adami RC, Alsante KM, Antipas AS, Arenson DR, Carrier R, Hong JY, Landis MS, Lombardo F, Shah JC, Shalaev E, Smith SW, Wang H (2002) Hydrolysis in pharmaceutical formulations. Pharm Dev Technol 7(2):113–146. doi:10.1081/PDT-120003494
5. Waterman KC, Adami RC, Alsante KM, Hong JY, Landis MS, Lombardo F, Roberts CJ (2002) Stabilization of pharmaceuticals to oxidative degradation. Pharm Dev Technol 7(1):1–32. doi:10.1081/PDT-120002237
6. Waterman KC, Adami RC (2005) Accelerated aging: prediction of chemical stability of pharmaceuticals. Int J Pharm 293(1–2):101–125. doi:10.1016/j.ijpharm.2004.12.013
7. Hochlowski J, Cheng XH, Sauer D, Djuric S (2003) Studies of the relative stability of TFA adducts vs non-TFA analogues for combinatorial chemistry library members in DMSO in a repository compound collection. J Comb Chem 5(4):345–349. doi:10.1021/cc0300107
8. Kozikowski BA, Burt TM, Tirey DA, Williams LE, Kuzmak BR, Stanton DT, Morand KL, Nelson SL (2003) The effect of freeze/thaw cycles on the stability of compounds in DMSO. J Biomol Screen 8(2):210–215. doi:10.1177/1087057103252618
9. Kozikowski BA, Burt TM, Tirey DA, Williams LE, Kuzmak BR, Stanton DT, Morand KL, Nelson SL (2003) The effect of room-temperature storage on the stability of compounds in DMSO. J Biomol Screen 8(2):205–209. doi:10.1177/1087057103252617
10. Engeloch C, Schopfer U, Muckenschnabel I, Le Goff F, Mees H, Boesch K, Popov M (2008) Stability of screening compounds in wet DMSO. J Biomol Screen 13(10):999–1006. doi:10.1177/1087057108326536
11. Popa-Burke I, Novick S, Lane CA, Hogan R, Torres-Saavedra P, Hardy B, Ray B, Lindsay M, Paulus I, Miller L (2014) The effect of initial purity on the stability of solutions in storage. J Biomol Screen 19(2):308–316. doi:10.1177/1087057113492201
12. Zitha-Bovens E, Maas P, Wife D, Tijhuis J, Hu QN, Kleinoder T, Gasteiger J (2009) COMDECOM: predicting the lifetime of screening compounds in DMSO solution. J Biomol Screen 14(5):557–565. doi:10.1177/1087057109336953
13. Xu J, Hagler A (2002) Chemoinformatics and drug discovery. Molecules 7(8):566–600. doi:10.3390/70800566
14. Cignitti M, Allen TL (1959) Bond energies and the interactions between next-nearest neighbors. I. Saturated hydrocarbons, diamond, sulfanes, S8, and organic sulfur compounds. J Chem Phys 43(12):4472–4478. doi:10.1021/ja00965a011
15. Berger JO (1993) Statistical decision theory and Bayesian analysis. Springer series in statistics, 2nd edn. Springer, New York
16. Xu J (1997) C-13 NMR spectral prediction by means of generalized atom center fragment method. Molecules 2(8):114–128. doi:10.3390/20800114
17. Kuhne R, Ebert RU, Schuurmann G (2009) Chemical domain of QSAR models from atom-centered fragments. J Chem Inf Model 49(12):2660–2669. doi:10.1021/ci900313u
18. Yan X, Gu Q, Lu F, Li J, Xu J (2012) GSA: a GPU-accelerated structure similarity algorithm and its application in progressive virtual screening. Mol Divers 16(4):759–769. doi:10.1007/s11030-012-9403-0
19. Klon AE (2009) Bayesian modeling in virtual high throughput screening. Comb Chem High Throughput Screen 12(5):469–483. doi:10.2174/138620709788489046
20. Chen L, Li YY, Zhao Q, Peng H, Hou TJ (2011) ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. Mol Pharm 8(3):889–900. doi:10.1021/mp100465q
21. Broccatelli P (2012) QSAR models for P-glycoprotein transport based on a highly consistent data set. J Chem Inf Model 52(9):2462–2470. doi:10.1021/ci3002809
22. Martins IF, Teixeira AL, Pinheiro L, Falcao AO (2012) A Bayesian approach to in silico blood–brain barrier penetration modeling. J Chem Inf Model 52(6):1686–1697. doi:10.1021/ci300124c
23. Townsend JA, Glen RC, Mussa HY (2012) Note on naive Bayes based on binary descriptors in Cheminformatics. J Chem Inf Model 52(10):2494–2500. doi:10.1021/ci200303m
24. Varnek A, Baskin I (2012) Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model 52(6):1413–1437. doi:10.1021/ci200409x
25. Sun HM (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. J Med Chem 48(12):4031–4039. doi:10.1021/jm050180t
26. Prathipati P, Ma NL, Keller TH (2008) Global Bayesian models for the prioritization of antitubercular agents. J Chem Inf Model 48(12):2362–2370. doi:10.1021/ci800143n
27. Xia XY, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. J Med Chem 47(18):4463–4470. doi:10.1021/jm0303195
28. Sun HM (2006) An accurate and interpretable Bayesian classification model for prediction of hERG liability. ChemMedChem 1(3):315–322. doi:10.1002/cmdc.200500047
29. Mussa HY, Mitchell JB, Glen RC (2013) Full "Laplacianised" posterior naive Bayesian algorithm. J Cheminform 5(1):37. doi:10.1186/1758-2946-5-37
30. Singh N, Chaudhury S, Liu R, AbdulHameed MD, Tawa G, Wallqvist A (2012) QSAR classification model for antibacterial compounds and its use in virtual screening. J Chem Inf Model 52(10):2559–2569. doi:10.1021/ci300336v
31. Xu J (2002) A new approach to finding natural chemical structure classes. J Med Chem 45(24):5311–5320. doi:10.1021/jm010520k
32. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Res 39:D1035–D1041. doi:10.1093/nar/gkq1126
33. Zhou Y, Zhou B, Chen K, Yan SF, King FJ, Jiang S, Winzeler EA (2007) Large-scale annotation of small-molecule libraries using public databases. J Chem Inf Model 47(4):1386–1394. doi:10.1021/ci700092v
34. Yan A, Hu X, Wang K, Sun J (2013) Discriminating of ATP competitive Src kinase inhibitors and decoys using self-organizing map and support vector machine. Mol Divers 17(1):75–83. doi:10.1007/s11030-012-9411-0
35. Rishton GM (1997) Reactive compounds and in vitro false positives in HTS. Drug Discov Today 2:382–384. doi:10.1016/S1359-6446(97)01083-0

36. Sushko I, Novotarskyi S, Korner R, Pandey AK, Cherkasov A, Lo JZ, Gramatica P, Hansen K, Schroeter T, Muller KR, Xi LL, Liu HX, Yao XJ, Oberg T, Hormozdiari F, Dao PH, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV (2010) Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model 50(12):2094–2111. doi:10.1021/ci100253r
37. Vorberg S, Tetko IV (2014) Modeling the biodegradability of chemical compounds using the online CHEmical modeling environment (OCHEM). Mol Inform 33(1):73–85. doi:10.1002/minf.201300030
38. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des 25(6):533–554. doi:10.1007/s10822-011-9440-2