Journal of Chemical Information and Modeling

Article

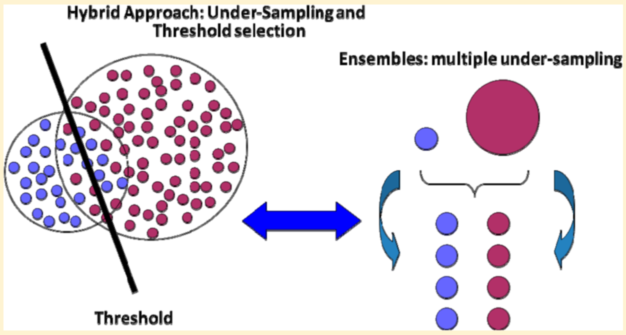
pubs.acs.org/jcim

QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem

Alexey V. Zakharov,[†] Megan L. Peach,[‡] Markus Sitzmann,[†,§] and Marc C. Nicklaus*[,†]

[†]CADD Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, DHHS, NCI-Frederick, 376 Boyles St., Frederick, Maryland 21702, United States

[‡]Basic Science Program, Leidos Biomedical, Inc., Computer-Aided Drug Design Group, Chemical Biology Laboratory, Frederick National Laboratory for Cancer Research, 376 Boyles St., Frederick, Maryland 21702, United States


Hybrid Approach: Under-Sampling and Threshold selection
Ensembles: multiple under-sampling
Threshold

ABSTRACT: Many of the structures in PubChem are annotated with activities determined in high-throughput screening (HTS) assays. Because of the nature of these assays, the activity data are typically strongly imbalanced, with a small number of active compounds contrasting with a very large number of inactive compounds. We have used several such imbalanced PubChem HTS assays to test and develop strategies to efficiently build robust QSAR models from imbalanced data sets. Different descriptor types [Quantitative Neighborhoods of Atoms (QNA) and "biological" descriptors] were used to generate a variety of QSAR models in the program GUSAR. The models obtained were compared using external test and validation sets. We also report on our efforts to incorporate the most predictive of our models in the publicly available NCI/CADD Group Web services (http://cactus.nci.nih.gov/chemical/apps/cap).

■ INTRODUCTION

PubChem is a very large public database of small molecules. It was designed as a public repository for compound structures and their biological properties. The bioactivity results in PubChem have been contributed by more than a hundred organizations. The majority of the data however come from the Molecular Libraries Probe Production Centers Network (MLPCN), in its previous version called the Molecular Libraries Screening Center Network (MLSCN), managed by the NIH Molecular Libraries Program (MLP).[1] This program aims to discover chemical probes through high-throughput screening (HTS) of small molecules to support chemical biology research. PubChem is organized as three linked databases: Substance, BioAssay, and Compound. The Substance database contains primarily structures supplied by depositors. The BioAssay database contains assay results for substances. The Compound database contains unique structures derived by structural standardization of the records in the Substance database.[2] As of the time of this writing, PubChem includes data on about 47 million unique compounds and results from more than 700,000 assays, with the total number of active compounds being 906,459 (as of July 2013).

ChEMBL[3] is another example of a public database of compound and activity data. ChEMBL is a database of bioactive drug-like molecules. It is supported by the European Bioinformatics Institute. The data in ChEMBL are abstracted and curated from primary scientific literature and include information about compound structures and their biological activities.

The comprehensive information on small molecules and their biological activities in PubChem and ChEMBL offers great opportunities for researchers in the fields of chemical biology, medicinal chemistry, and chemoinformatics.[4] The PubChem and ChEMBL databases have different distributions of compounds as to their activities in each assay, resulting from the different nature of data acquisition in each of these two resources. PubChem incorporates HTS data, which are highly imbalanced in general, that is, they have a small ratio of active compounds to inactive ones. This data distribution corresponds to what would be found in a completely unbiased approach of a true random selection of screening compounds from all of chemistry space, which could be termed the "natural" distribution. In this "natural" case, the number of active compounds would be much smaller than the number of inactive compounds for each particular activity. ChEMBL data, in contrast, are in most cases extracted from literature; thus they are more balanced but by the same token more biased toward active compounds. This is not surprising given that there are strong incentives to publish positive results (i.e., active compounds) and only limited rewards for publishing negative ones (i.e., inactive molecules); although for model-building, the latter results are as important as the former. Thus, one could see the data distribution in ChEMBL as more artificial and

Received: December 11, 2013
Published: February 13, 2014

distinct from "natural" biological measurements. However, the advantage of this data distribution is that it ranges up to high potency values of the included compounds. The PubChem database, conversely, has many compounds with low potency. According to Li et al.,[1] approximately 40% of the targets in PubChem had no active compound with potency better than 10 μM. We therefore see it as an important challenge to extract maximum predictivity from PubChem data, which may benefit especially the academic part of the community of scientists working in drug development that may have to rely to a larger extent on public data than their industrial colleagues. This project is an attempt to meet this challenge by applying our computational methodology to imbalanced data sets in PubChem. (A complementary type of imbalance on the side of the independent variables, that is molecular descriptors, exemplified by issues such as the diversity of the training and test sets, coverage of chemical space, singletons versus smaller or larger series of closely related analogs, etc., while equally important for model building, is not the topic of this study.)

Whether data sets are balanced or imbalanced, it is often desirable to be able to use these data to model compound potency against a given target for additional molecules, which is the realm of quantitative structure−activity relationships (QSAR). The main idea of all QSAR methods is to describe relationships between activity measures and compound structural descriptors and to create a model that can be used to predict the activity of new compounds. Many different techniques have been used for this task. Among the most popular ones at present are Random Forests,[5] Support Vector Machines[6] (SVM), and Artificial Neural Networks.[7] Unfortunately, most popular machine learning approaches have shown weak performance with imbalanced data. The reason for this is that they are based on the fundamental premise that all data points, that is, elements of the training set, have the same a priori importance with the consequence that the majority class has the tendency to swamp out the signal from the minority class. As a result, researchers have tried to address this problem by a whole slew of approaches.

One of the decisions one has to make early on in the process is whether to use probabilistic approaches such as the Naïve Bayes classifier, which works quite well with highly imbalanced data. But it is also well known[8,9] that probabilistic approaches provide poorer prediction accuracy in comparison with modern machine learning techniques such as Random Forests or SVM, which led us to focus on these types of techniques and their possible improvement for imbalanced data sets. Excluding probabilistic approaches, these machine learning techniques can be divided into algorithm-based methods and data-based methods.

Algorithm-based methods deal with cost-sensitive learning and use penalties for misclassifying the minority class, which in the case of the PubChem data sets would be the small subset of active molecules. Some authors have provided specific modifications to machine learning approaches. For instance, Li et al. proposed a modification of SVM, called the GSVM-RU method,[4] which extracts informative inactive samples and uses them together with all active samples for the construction of support vectors. Chen et al.[10] proposed an algorithm called Weighted Random Forest, which assigns a weight to each class with the minority class given a larger weight. Chang et al.[11] proposed a similar modification to the SVM method, which assigns a weight to each class, and implemented it in the LiBSVM program.[12] Another modification of SVM is based on optimizing the performance measures such as the ROC area and has been implemented in SVMPerf[13] by Joachims.[14] The drawback of algorithm-based methods is that they require algorithm-specific modifications. It is worth pointing out that many algorithm-based modifications have been described in the literature, but most of them have not been implemented in readily available software and thus are not directly accessible to the medicinal chemist or chemoinformatician.

Data-based methods deal with the sampling technique and therefore can be used independently of any specific machine learning method. There are two types of sampling methods: under-sampling and over-sampling. Single over-sampling is used for the generation of new synthetic minority class members by interpolating between several examples. Chawla et al. proposed the SMOTE method (synthetic minority over-sampling technique) and successfully used it for modeling of imbalanced data sets.[15] Another sampling method is under-sampling. The single under-sampling method is used for reducing the number of samples in the majority class to make it equivalent in size to the minority class. This method has shown good performance in several publications. For instance, Chen et al.[16] used the under-sampling method for toxicity modeling of *Tetrahymena pyriformis*. Sun et al. applied the same method to the prediction of cytochrome P450 profiles of environmental chemicals. Newby et al.[17] modeled imbalanced oral absorption data sets. Chen et al. compared the over-sampling approach with under-sampling and showed that the under-sampling method performed more consistently. In other work, Drummond and Holte reached the same conclusion.[18] The advantage of sampling-based approaches is that they are independent of the specific machine learning method used and thus can be applied to any classification algorithm. However, the simple under-sampling method reduces chemistry space in the majority class, which may be the reason for its decrease in accuracy. To avoid this, some authors have used multiple under-sampling methods (ensembles), which generate different bootstrap samples of equal class size in the training set to build ensemble models. For instance, Kondratovich et al.[19] used multiple under-sampling methods for prediction of the assignment of organic compounds to different pharmacological groups. Other authors have proposed using a rational reduction of samples from the majority class by similarity methods.[20] For more information, the reader is referred to the excellent review article by Varnek and Baskin[21] discussing the problem of imbalanced data in chemoinformatics. To our knowledge there is however no comprehensive comparison of these different approaches that would analyze their ability to model imbalanced data. We are trying to address this lack of comparison of methods to some extent by analyzing several common strategies of imbalanced data modeling in this work. We then propose a new hybrid method, which includes both cost-sensitive learning and under-sampling approaches. We compare all methods on five different HTS data sets extracted from PubChem. Finally, we are reporting on our efforts to make all QSAR models developed in this work freely available in our Chemical Activity Predictor Web service: http://cactus.nci.nih.gov/chemical/apps/cap.

■ MATERIALS AND METHODS

Data Sets. Five confirmatory bioassay data sets in PubChem were used for the construction of modeling sets: AID 504466, AID 485314, AID 485341, AID 624202, and AID 651820. All these quantitative high-throughput screening (qHTS) data had

Table 1. Characteristics of PubChem HTS Assays Used for QSAR Modeling

AID	name	initial number	after preprocessing	active	inactive	ratio
504466	genotoxicity inductors in HEK293T cells	330,115	310,403	4108	306,295	1:75
485314	DNA polymerase beta inhibitors	334,467	306,830	4348	302,482	1:70
485341	AmpC beta-lactamase inhibitors	330,683	285,970	1694	284,276	1:168
624202	BRCA1 activators	376,014	351,201	3902	347,299	1:89
651820	hepatitis C virus inhibitors	339,561	268,119	10,727	257,392	1:24

been collected at the NIH Chemical Genomics Center (NCGC), now part of the NIH National Center for Advancing Translational Sciences (NCATS), with each assay having been run on approximately 300,000 screening samples.

(1) AID 504466[22] is a qHTS screen for small molecules that induce genotoxicity in human embryonic kidney cells (HEK293T) expressing luciferase-tagged ELG1. This assay was developed to find promising inhibitors of DNA replication, which could be potential anticancer agents.

(2) AID 485314[23] is a qHTS assay for inhibitors of DNA polymerase beta. DNA polymerase beta plays an important role in the repair system of human cells and is thus a promising target for therapeutic modulation of the response to radiation treatment and DNA-damaging drugs.

(3) AID 485341[24] is a qHTS assay for inhibitors of AmpC beta-lactamase. It was one of a series of assays conducted to distinguish aggregators versus nonaggregators by adding, or not adding, detergent. The assay used here was run without detergent. Compounds that inhibit only in the absence of detergent are considered likely promiscuous aggregators.

(4) AID 624202[25] is a qHTS assay to identify small molecule activators of BRCA1 expression. BRCA1 has been implicated in a wide array of cellular activities, including DNA damage repair, cell-cycle checkpoint control, growth inhibition, apoptosis, transcriptional regulation, chromatin remodeling, protein ubiquitylation, and mammary stem cell self-renewal and differentiation. Increase in BRCA1 expression would enable cellular differentiation and restore tumor suppressor function, resulting in delayed tumor growth and less aggressive, more treatable breast cancers. Promising activators of BRCA1 expression could be novel preventative or therapeutic agents against breast cancer.

(5) AID 651820[26] is a qHTS assay for inhibitors of hepatitis C virus (HCV). This assay was developed to find novel HCV inhibitors as new therapies for hepatitis C.

For each set of assay data, preprocessing was performed. All compounds with inconclusive results were removed. All structures were normalized using the CACTVS chemoinformatics toolkit.[27] All salts and mixtures were removed. The total number of compounds remaining after preprocessing for each assay, along with the activity distribution and ratio of inactives to actives, are given in Table 1.

The most balanced ratio of active to inactive compounds was found for the assay for inhibitors of hepatitis C virus (1:24). The most imbalanced ratio was found for the assay for inhibitors of AmpC beta-lactamase (1:168). The average ratio was about 1:70. These different ratios of active to inactive compounds give us the opportunity to compare by level of imbalance the various approaches for dealing with imbalanced data.

For the modeling, each data set from each assay after preprocessing was randomly divided in an 80:20 split into the training set used to create the QSAR models and a test set used to assess their external predictive accuracy. Thus, the number of compounds in each test set was at least 55,000 compounds, which assured us that we would obtain statistically significant results.

■ METHODS

General Approach and Software Used. Seven different imbalanced learning methods were applied to each training set (see below). Several QSAR models were created for each method. All QSAR models were built using the program GUSAR (General Unrestricted Structure Activity Relationships; version 2013).[28] For model construction, GUSAR uses Quantitative Neighborhoods of Atoms (QNA) descriptors[28,29] and "biological" descriptors (PASS-based predictions)[30,31] and applies a self-consistent regression (SCR) algorithm.[30,28] The QSAR models developed for each imbalanced learning method were validated on the corresponding test sets.

Imbalanced Learning Methods. *(1). One-Sided Random Sampling.* Under-sampling was done by randomly selecting compounds from the majority class, which in this case are the inactive compounds, until the total number of selected inactive compounds was equal to the number of active compounds in the minority class. As a result, the training set is represented by one data set, which includes an equal number of active and inactive compounds.

(2). Multiple Under-Sampling. The majority class of the training set was randomly sampled up to the size of the minority class. This procedure was repeated multiple times until all compounds from the majority class had become part of a training set at least once. A compound is removed from the majority class once it has been included in a training set. As a result, many training sets were constructed, each including all active compounds and the same number of inactive compounds selected randomly. Thus, the number of training sets for each particular assay/activity corresponds to the value of the ratio of inactives to actives in this assay. The advantage of this method is that the chemistry coverage of inactive compounds is much broader in comparison to the one-sided sampling method. The drawback is the larger number of training sets that are formed.

(3). Under-Sampling Clusterization. Under-sampling was done by selecting compounds from the majority class using clustering techniques until the total number of inactive compounds was equal to the number of active compounds in the minority class. This was done by clustering the compounds in the majority class with the number of clusters preset to the number of compounds in the minority class. The central compound from each cluster was extracted to form a new balanced training set. As a result, the training set is represented by one data set that includes an equal number of active and inactive compounds. To perform the clustering, we used the program Pipeline Pilot[32], version 5.0, employing the "Cluster Molecules" component and FCFP_4 fingerprints as structure descriptors. The advantage of this method is that it generates a chemical distribution of inactive compounds that is similar to the majority class as a whole.

(4). Under-Sampling Diversity. Under-sampling was done by selecting compounds from the majority class using diversity analysis until the total number of selected inactive compounds was equal to the number of active compounds in the minority class. To this goal, compounds from the majority class were analyzed using the maximum dissimilarity method as implemented in the "Diverse Molecules" component of Pipeline Pilot, version 5.0. FCFP_4 fingerprints were used as structure descriptors. The most diverse compounds were extracted from the majority class and were used to form a new balanced training set, which includes an equal number of active and inactive compounds. The advantage of this method is that it retains the chemical diversity of the majority class in the inactive training set compounds.

(5). Under-Sampling Similarity. Under-sampling was done by selecting compounds from the majority class using similarity analysis. The similarity between compounds from the minority and majority classes was calculated using Multilevel Neighborhoods of Atoms (MNA) descriptors[33] and Tanimoto coefficients. According to the computed similarity values, the compounds in the majority class that were most similar to those in the minority class were extracted and were used to form a new balanced training set containing an equal number of active and inactive compounds. The advantage of this method is that it allows one to easily separate compounds in the training set from the remaining compounds of the majority class based on similarity. The drawback is that it is difficult for machine learning approaches to discriminate active versus inactive compounds when the inactives are very similar to the actives,[20] which is typically the outcome of this method.

(6). Adjusting Decision Threshold Approach. The main idea of this approach is to adjust the decision threshold (boundary) in assigning class memberships. This can be done in different ways. Here, we used the simplest one. We plot the curve of both sensitivity and specificity values as a function of different cutoff values (thresholds). The point where the sensitivity and specificity curves intersect determines the decision threshold. This plot can be constructed using external K-fold cross-validation or leave-one-out cross-validation (LOO−CV) procedures. The threshold after the validation procedure would be adjusted in accordance with the nature of the data distribution (degree of imbalancedness). This type of approach to handle imbalanced training sets is data set independent, that is, the number of compounds in the training set remains unchanged, and thus, the whole set remains imbalanced.

(7). Hybrid Method: Under-Sampling and Threshold Approach. We propose this method as a novel approach to combine the advantages of the decision threshold approach with the under-sampling approach. Here, under-sampling was done by randomly selecting compounds from the majority class until the total number of inactive compounds was three times larger than the number of active compounds in the minority class. Thus, we are retaining an imbalanced distribution of inactive compounds but are reducing the ratio to 1:3. In addition, for each assay's compound set, the decision threshold was calculated using a leave-one-out cross-validation procedure in the model development.

Descriptors. For the development of our QSAR models, two types of descriptors, QNA descriptors[28,34] and "biological" descriptors, based on the PASS algorithm[30,31] were used in combination with 10 whole-molecule descriptors. All these descriptors have been implemented in GUSAR.

The calculation of QNA descriptors is based on the connectivity matrix (C) of a given molecule and also on the standard values of the ionization potential (IP) and the electron affinity (EA) for each atom in the molecule.

For any given atom i in the molecule, the QNA descriptors are calculated as

$$P_i = B_i \sum_k \left(\exp\left(-\frac{1}{2}C\right)\right)_{ik} B_k$$

$$Q_i = B_i \sum_k \left(\exp\left(-\frac{1}{2}C\right)\right)_{ik} B_k A_k$$

where the k are all other atoms in the molecule and

$$A_k = \frac{1}{2}(\mathrm{IP}_k + \mathrm{EA}_k), \quad B_k = (\mathrm{IP}_k - \mathrm{EA}_k)^{-1/2}$$

The P and Q values are calculated for all atoms of the molecule. Two-dimensional Chebyshev polynomials are used for approximating the functions P and Q. Thus, the independent variables are calculated as average values of particular two-dimensional Chebyshev polynomials of the P and Q values for all atoms in the molecule.

In addition, GUSAR allows the creation of QSAR models based on predicted biological activity profiles of compounds. This is done by running the PASS algorithm[30,31] on each compound represented as a list of MNA descriptors[33] to predict the compound's biological activity profile. The current version of PASS (version 12) predicts 6400 types of biological activity with a mean prediction accuracy of about 95%. The list of predicted biological activities includes pharmacotherapeutic effects, mechanisms of action, adverse and toxic effects, metabolic terms, susceptibility to transporter proteins, and activities related to gene expression. The results of the PASS procedure are output as a list of the difference between the probabilities, for each biological activity, that the compound is active (P_a) or inactive (P_i), respectively. For building the different QSAR models in GUSAR, subsets of these P_a−P_i values were randomly selected from the total list of predicted biological activities as input independent variables for the regression analysis.

GUSAR also allows the calculation of whole-molecule descriptors: topological length, topological volume, lipophilicity, number of positive charges, number of negative charges, number of hydrogen bond donors, number of hydrogen bond acceptors, number of aromatic atoms, molecular weight, and number of halogen atoms. These descriptors were used in combination with the QNA and "biological" descriptors described above.

Self-Consistent Regression. For generating QSAR models, GUSAR uses a self-consistent regression algorithm. SCR is based on the regularized least-squares method. The basic purpose of the SCR method is to remove the variables that poorly describe the modeled value but to retain the set of variables correctly representing the existing relationship. It has been shown[28,35] that the self-consistent regression realized in GUSAR can be successfully applied to various QSAR tasks and that the prediction results achieved with GUSAR were comparable to or better than those obtained by other QSAR methods based on different machine learning approaches. It therefore seemed a natural choice to use this method for the investigation of the imbalanced learning problems. The details of the algorithms for the descriptor calculation and self-

Table 2. Test Set Prediction Quality Parameters for Each Imbalanced Learning Approach

	multiple under-sampling		threshold, ratio 1:3 under-sampling		one-sided under-sampling		similarity under-sampling		cluster under-sampling		diversity under-sampling		only threshold selection	
assay AID	BA	GM	BA	GM	BA	GM	BA	GM	BA	GM	BA	GM	BA	GM
504466	0.84	0.84	0.83	0.83	0.80	0.80	0.76	0.76	0.82	0.82	0.72	0.72	0.77	0.75
485314	0.78	0.78	0.78	0.78	0.76	0.76	0.67	0.67	0.69	0.68	0.68	0.65	0.72	0.69
485341	0.70	0.70	0.69	0.69	0.66	0.66	0.52	0.52	0.61	0.58	0.58	0.51	0.50	0.06
624202	0.80	0.80	0.79	0.79	0.76	0.76	0.67	0.67	0.76	0.75	0.72	0.69	0.60	0.47
651820	0.78	0.78	0.77	0.77	0.75	0.75	0.70	0.70	0.71	0.70	0.69	0.68	0.75	0.74
average	0.78	0.78	0.77	0.77	0.74	0.74	0.66	0.66	0.72	0.71	0.68	0.65	0.67	0.54

BA: balanced accuracy. GM: geometric mean (G-mean).

consistent regression methods have been described previously.[28,34]

Applicability Domain. GUSAR uses three different approaches for estimation of the applicability domain of each model: similarity, leverage, and accuracy assessment.[35]

Similarity. The three nearest neighbors from the training set are calculated for each compound under study using an estimation of similarity. The pairwise similarity of the compound with each of its three neighbors is estimated as Pearson's coefficient calculated in the space of the independent variables obtained after SCR. The average of these three similarity values is used as the applicability domain (AD) of the model. In this study, an AD threshold of 0.7 was used.

Leverage. The leverage value is also used for domain applicability assessment. The leverage values represent the "distance" of each molecule to the model's structural space and are a measure of the contribution of the n^{th} molecule to its own predicted value. Thus, they can be used to identify outliers and are calculated as

$$\text{Leverage} = x^{\text{T}}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}x$$

where x is the vector of the descriptors of a query compound, and $\mathbf{X}$ is the matrix formed with rows corresponding to the descriptors of the molecules from the training set. The leverage warning value was calculated for each compound in the training set, and then the distribution of these values was determined. In this study, the warning level for leverage values was set to the 99th percentile, that is, if a compound from the external test set had a leverage value exceeding this warning level, then this compound was considered as being outside the applicability domain.

Accuracy Assessment. For this type of assessment of the applicability domain, the following equation is used

$$\text{AD}_{\text{value}} = \text{RMSE}_{\text{3NN}}/\text{RMSE}_{\text{train}}$$

where AD_{value} is the applicability domain value, RMSE_{3NN} is the root-mean-square error of prediction of the three most similar compounds from the training set (as with the Similarity method), and $\text{RMSE}_{\text{train}}$ is the root-mean-square error of predictions for the training set.

In this study, a threshold of 1.0 was used for the AD calculated by accuracy assessment.

Consensus Modeling. GUSAR allows the creation of different QSAR models for each activity/end-point based on different types of descriptors (QNA descriptors and "biological" descriptors, see above). The final predicted value for each activity/end-point is estimated by including a weighted average of the predicted values from the set of QSAR models (for predictions that are within their respective applicability domains). The value obtained from each model is weighted by the similarity value calculated in the estimation of its applicability domain.

Evaluation of Prediction Accuracy. For estimating the accuracy of prediction, the following statistical parameters were calculated:

(1) Sensitivity: probability of predicting "positive" (active) when the true outcome is positive.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

where TP is true positive, and FN is false negative.

(2) Specificity: probability of predicting "negative" (inactive) when true outcome is negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TN is true negative, and FP is false positive.

(3) Balanced Accuracy: balance between Sensitivity and Specificity.

$$\text{Balanced Accuracy} = (\text{Sensitivity} + \text{Specificity})/2$$

(4) G-mean: geometric mean of Sensitivity and Specificity values

$$\text{G-mean} = (\text{Sensitivity} \times \text{Specificity})^{1/2}$$

■ RESULTS

QSAR Modeling. For each training set obtained from the rebalancing procedures except for the multiple under-sampling method, 10 models based on "biological" descriptors and 10 models based on QNA descriptors were created. Thus, 20 QSAR models were created for each training set represented by one data set.

With the multiple under-sampling procedure, many training sets were obtained for each particular assay. For each of these training sets, two QSAR models were constructed: one based on "biological" descriptors and one based on QNA descriptors. The total number of models generated for each assay thus computes as twice the value of the ratio of inactives to actives. For example, assay AID 504466 has a ratio of 1:75, so the number of models created was 150.

To select the most predictive models obtained by GUSAR, a leave-10%-out cross-validation procedure was performed 10 times for each model. From the full set of generated models, we selected only those that satisfied the following conditions: values of balanced accuracy for LOO−CV and the leave-10%-out cross-validation procedures had to exceed 0.6. The selected models were used for consensus predictions of the external test

Table 3. Test Set Predictions Quality Parameters for Combined Imbalanced Learning Approaches

	combination hybrid and cluster under-sampling		hybrid under-sampling		cluster under-sampling	
assay AID	BA	GM	BA	GM	BA	GM
504466	0.85	0.85	0.83	0.83	0.82	0.82
485314	0.74	0.73	0.78	0.78	0.69	0.68
485341	0.67	0.66	0.69	0.69	0.61	0.58
624202	0.80	0.80	0.79	0.79	0.76	0.75
651820	0.75	0.75	0.77	0.77	0.71	0.70
average	0.76	0.76	0.77	0.77	0.72	0.71

BA − balanced accuracy, GM − geometric mean (G-mean).

set for each activity/end-point, taking into account the applicability domain of these models.

Comparison of Imbalanced Learning Methods. The seven different imbalanced learning approaches described above were applied to modeling of the five HTS assay data sets. Balanced accuracy values calculated for each test set for each of the imbalanced learning approaches are shown in Table 2.

Table 2 shows that the best results in terms of the balanced accuracy and G-mean evaluation values were obtained for the multiple under-sampling method as well as our hybrid method. A bootstrap method with 10,000 repeats showed that these results were better than the results obtained by other sampling methods with a statistical significance at the p = 0.001 level. The average accuracy of prediction achieved for the five test sets exceeded 0.75 for both methods. Poorer results were obtained with the similarity, diversity under-sampling, and simple threshold selection methods. The clusterization and one-sided under-sampling methods produced results in the middle of the range with an accuracy of prediction around 0.70.

The reason for the poor results achieved by the similarity and diversity methods is perhaps due to the artificial nature of chemical (sub)space that was produced by these methods. Indeed, a distribution of chemical compounds restricted by similarity in the first case and by diversity in the second case conceivably does not well represent the chemical space of the test sets, which were kept untouched. Indirectly, this can be proven by the cluster under-sampling approach, which showed better results in comparison to both the similarity and diversity methods. This method retains the initial chemical space in the training set due to data clusterization and does not create an artificial distribution as the similarity and diversity methods do. The same explanation might be applied to the one-sided under sampling approach, which due to the random extraction of compounds from the initial pool also more likely preserved the initial chemical space.

It is no surprise then that the multiple under-sampling approach showed the best results. In comparison with the other methods, this approach allows all compounds from the majority class to be retained for model development. Thus, this method provides a better chemical coverage of the test set compounds. In addition, due to the nature of the method, all models are created from balanced data sets.

Our new hybrid method gave the second best results. The reason for this is that the hybrid method has better chemical coverage than the one-sided approach, and it also retains more of the chemical space in the training set in comparison to the similarity and diversity under-sampling approaches. In addition, the application of threshold selection used in this method produces balanced prediction results.

All compounds from the majority class were used for model development in the adjusting decision threshold approach. However, this method gave poor results in terms of the G-mean values for two of the five training sets. Analysis of these results suggests that this method cannot be successfully used for extremely imbalanced data sets with a ratio of actives to inactives more imbalanced than 1:80 but may still be applied to smaller imbalanced data sets.

Combination of Imbalanced Learning Methods. Taking into account the pros and cons of each imbalanced learning approach, we analyzed various two-method combinations. For this purpose, models obtained from the hybrid and cluster imbalanced methods were combined, and prediction results were aggregated to give consensus predictions. Both of these methods retain the initial chemical space in different ways: random selection and clusterization. Thus, we reasoned that the combination of both of them might produce improved results. The best approach, multiple under-sampling, was excluded from any combination because this method includes all the compounds from the training set to begin with. Results achieved by the combined method and by two separate hybrid and cluster methods were compared as to their respective balanced accuracy and G-mean values (Table 3).

Table 3 shows that combined methods produce better results compared to two separate methods in two of the five assays (504466, 624202). Analysis of these results shows that for these two assays the difference between the accuracy values of the two separate methods is small: 0.01 for assay 504466 and 0.03 for assay 624202. For the remaining three assays, the difference in accuracy values between the two separate methods exceeds 0.06. Thus, our observations suggest that an improvement in the results by combining two imbalanced methods can be achieved in situations when the accuracies of the two separate methods are close to each other.

External Validation of Imbalanced Methods. During the course of this study, a new data set of genotoxicity inductors in human embryonic kidney cells became available from NIH NCATS. These new confirmatory bioassay data are presented in PubChem as AID 651632. According to the assay description in PubChem, compounds in this bioassay data set were assayed under the same conditions as used in AID 504466. Thus, this provided us with an additional opportunity for external validation of our QSAR models developed for assay AID 504466 using the various imbalanced learning approaches.

The initial number of compounds in AID 651632 was 10,496. We preprocessed the data according to the procedure described above. The total number of compounds left in the validation set was 9402: 218 actives and 9184 inactives (ratio 1:42). Predictions for this set were calculated by the QSAR models obtained with three imbalanced learning approaches:

Table 4. External Validation of Imbalanced Methods

	combination of hybrid and cluster under-sampling		hybrid under-sampling		multiple under-sampling	
assay AID	BA	GM	BA	GM	BA	GM
651632	0.68 (AD: 100%)	0.62 (AD: 100%)	0.68 (AD: 100%)	0.65 (AD: 100%)	0.68 (AD: 100%)	0.66 (AD: 100%)
651632 (combined AD)	0.68 (AD: 95.2%)	0.63 (AD: 95.2%)	0.69 (AD: 79.3%)	0.66 (AD: 79.3%)	0.69 (AD: 92.2%)	0.67 (AD: 92.2%)

AD: applicability domain. BA: balanced accuracy. GM: geometric mean (G-mean). Lower row: predictions limited to those that fall in the combined applicability domains of the hybrid and cluster combination methods.

multiple under-samplings, hybrid under-sampling, and our combined method (hybrid and cluster under-sampling). The balanced accuracy and G-mean values calculated for this external validation set are shown in Table 4.

Predictions were made both with and without taking into account the combined applicability domains of the models (Table 4, lower and upper row, respectively), the former obviously corresponding to 100% AD coverage. All imbalanced learning methods yielded a similar accuracy of prediction calculated for full coverage of the external validation set (AD: 100% coverage), which generally exceeded 0.65. Limiting the prediction to those that fell in the combined applicability domains of the hybrid and cluster combination methods (bottom row in Table 4) yielded a slight increase in the coverage for the external validation set but did not improve the accuracy of prediction. The simple hybrid method produced similar prediction results as the multiple under-sampling approach but with less coverage. All methods showed a lower accuracy of prediction in comparison to the results obtained for the test set from AID 504466. The reasons for this may be that this validation set covers new chemical space and/or has a different distribution of active and inactive compounds. Indeed, the test set from AID 504466 has a ratio of 1:75, whereas the validation set from AID 615632 has a ratio of 1:42. Nevertheless, the best results achieved with the hybrid and multiple under-sampling methods for AID 615632 were closer to 0.70, which is quite good for an external validation test.

Chemical Activity Predictor Web Service. We have made the QSAR models developed with the GUSAR program utilizing the hybrid imbalanced learning approach freely available in our online Chemical Activity Predictor service at: http://cactus.nci.nih.gov/chemical/apps/cap. This service provides the user with two different ways of inputting chemical structures. The first one is a classical online chemical editor, which allows the desired structure to be drawn and submitted. The second one is based on our NCI/CADD Chemical Identifier Resolver[36] technology and allows the input of different types of structure identifiers: InChIKey, drug names, SMILES, IUPAC names, etc. This service allows the user to input several structures simultaneously. As output, the service provides binary prediction results for the five HTS assays for each compound. In addition, our service estimates the applicability domain of each QSAR model with the result that for each compound each prediction is annotated with either “In AD” or “Out of AD”, indicating whether one can be confident in the prediction or not. Performance tests on the current hardware showed that our service operates at a reasonable computational speed (approximately 5 compounds per second for the simultaneous prediction of five activities).

■ CONCLUSIONS

We have analyzed several common strategies for imbalanced data modeling. In total, seven methods were compared using HTS data sets extracted from PubChem. Our analysis led us to propose a new hybrid method, which includes both cost-sensitive learning and under-sampling approaches. We showed that the multiple under-sampling approach and our hybrid method provide more accurate prediction results than other methods. Our QSAR models showed a generally high accuracy of prediction, on average exceeding 0.75. In addition, the combination of two imbalanced learning approaches was investigated, with the result that some improvements could be achieved using this approach. We hope that our QSAR models may be useful for in silico screening of compound libraries for the five activities from the PubChem HTS bioassay collection. Our freely available Chemical Activity Predictor Web service provides public access to these QSAR models and may be found useful by researchers trying to find drug-like leads with desirable properties. In addition, this service allows the optimization of compounds across different activities simultaneously.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mn1@helix.nih.gov. Telephone: +1-301-846-5903.

Present Address

§Marcus Sitzmann: FIZ Karlsruhe − Leibniz-Institut für Informationsinfrastruktur, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project has been funded in part with federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under Contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States government.

■ ABBREVIATIONS

GUSAR, general unrestricted structure−activity relationships; PASS, prediction of activity spectra for substances; SCR, self-consistent regression; QSAR, quantitative structure−activity relationships; MNA, multilevel neighborhoods of atoms; QNA, quantitative neighborhoods of atoms

REFERENCES

(1) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052−1057.

(2) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Vol. *4*, Chapter 12, pp 217−241.

(3) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.

(4) Li, Q.; Wang, Y.; Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **2009**, *25*, 3310−3316.

(5) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(6) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(7) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Neural Networks in Building QSAR Models. In *Artificial Neural Networks: Methods and Applications*; Livingstone, D. J., Ed.; Methods in Molecular Biology Series, Vol. 458; Springer: Clifton NJ, 2008; pp 137−158

(8) Caruana, R.; Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics. In *ICML '06 Proceedings of the 23rd International Conference on Machine Learning*; Pittsburgh, PA, 2005, ACM Press: New York, 2005, pp 161−168.

(9) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of random forest and pipeline pilot naïve bayes in prospective QSAR predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.

(10) Chen, C.; Liaw, A.; Breiman, L. *Using Random Forest To Learn Imbalanced Data*; Statistics Department, University of California, Berkeley, Berkeley, CA, 2004.

(11) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1−27:27.

(12) LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed February 4, 2014).

(13) SVM-perf: Support Vector Machine for Multivariate Performance Measures. http://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html (accessed February 4, 2014).

(14) Joachims, T. A Support Vector Method for Multivariate Performance Measures. In *ICML '05 Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005, ACM Press: New York, 2005; pp 377−384.

(15) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.

(16) Chen, J.; Tang, Y. Y.; Fang, B.; Guo, C. In silico prediction of toxic action mechanisms of phenols for imbalanced data with random forest learner. *J. Mol. Graph. Modell.* **2012**, *35*, 21−27.

(17) Newby, D.; Freitas, A. A.; Ghafourian, T. Coping with unbalanced class data sets in oral absorption models. *J. Chem. Inf. Model.* **2013**, *53*, 461−474.

(18) Drummond, C.; Holte, R. C. *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling*; Workshop on Learning from Imbalanced Datasets II, International Council for Machinery Lubrication (ICML): Washington DC, 2003; pp 1−8.

(19) Kondratovich, E. P.; Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Fragmental descriptors in (Q)SAR: Prediction of the assignment of organic compounds to pharmacological groups using the support vector machine approach. *Russ. Chem. Bull.* **2009**, *58*, 657−662.

(20) Zhang, L.; Fourches, D.; Sedykh, A.; Zhu, H.; Golbraikh, A.; Ekins, S.; Clark, J.; Connelly, M. C.; Sigal, M.; Hodges, D.; Guiguemde, A.; Guy, R. K.; Tropsha, A. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J. Chem. Inf. Model.* **2013**, *53*, 475−492.

(21) Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413−1437.

(22) AID 504466, PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=504466 (accessed October 17, 2013).

(23) AID 485314, PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=485314 (accessed October 17, 2013).

(24) AID 485341, PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=485341 (accessed October 17, 2013).

(25) AID 624202, PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=624202 (accessed October 17, 2013).

(26) AID 651820, PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=651820 (accessed October 17, 2013).

(27) Ihlenfeldt, W.-D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(28) Filimonov, D. A.; Zakharov, A. V.; Lagunin, A. A.; Poroikov, V. V. QNA-based "Star Track" QSAR approach. *SAR QSAR Environ. Res.* **2009**, *20*, 679−709.

(29) Kokurkina, G. V.; Dutov, M. D.; Shevelev, S. A.; Popkov, S. V.; Zakharov, A. V.; Poroikov, V. V. Synthesis, antifungal activity and QSAR study of 2-arylhydroxynitroindoles. *Eur. J. Med. Chem.* **2011**, *46*, 4374−4382.

(30) Lagunin, A.; Zakharov, A.; Filimonov, D.; Poroikov, V. QSAR modelling of rat acute toxicity on the basis of PASS prediction. *Mol. Inform.* **2011**, *30*, 241−250.

(31) Zakharov, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Quantitative prediction of antitarget interaction profiles for chemical compounds. *Chem. Res. Toxicol.* **2012**, *25*, 2378−2385.

(32) Warr, W. A. Scientific workflow systems: Pipeline pilot and KNIME. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 801−804.

(33) Geronikaki, A.; Druzhilovsky, D.; Zakharov, A.; Poroikov, V. Computer-aided prediction for medicinal chemistry via the Internet. *SAR QSAR Environ. Res.* **2008**, *19*, 27−38.

(34) Lagunin, A. A.; Zakharov, A. V.; Filimonov, D. A.; Poroikov, V. V. A new approach to QSAR modelling of acute toxicity. *SAR QSAR Environ. Res.* **2007**, *18*, 285−298.

(35) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Filippov, I. V.; McCartney, H. J.; Smith, L. H.; Pugliese, A.; Nicklaus, M. C. Computational tools and resources for metabolism-related property predictions. 2. Application to prediction of half-life time in human liver microsomes. *Future Med. Chem.* **2012**, *4*, 1933−1944.

(36) NCI/CADD Chemical Identifier Resolver. http://cactus.nci.nih.gov/chemical/structure (accessed October 18, 2013).