Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-019-00255-3

Undersampling: case studies of flaviviral inhibitory activities

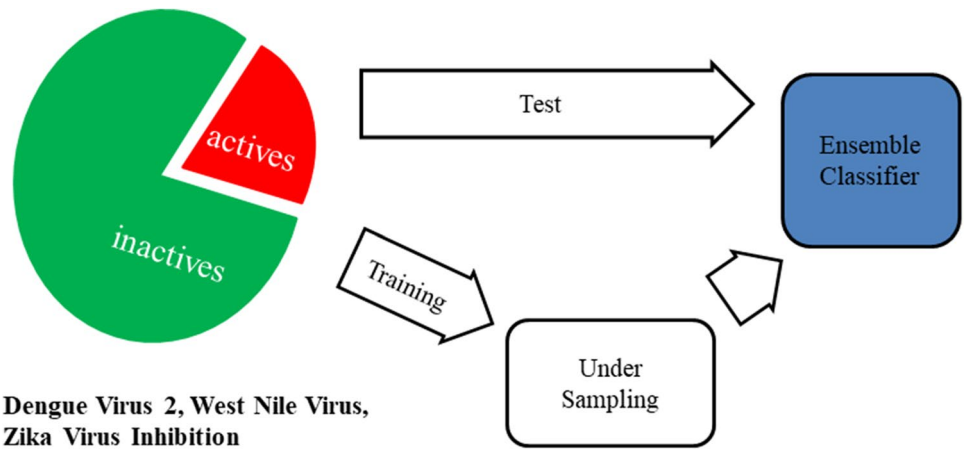
Stephen J. Barigye[1] · **José Manuel García de la Vega**[1] · **Juan A. Castillo-Garit**[2]

Received: 17 September 2019 / Accepted: 19 November 2019

Abstract

Imbalanced datasets, comprising of more inactive compounds relative to the active ones, are a common challenge in ligand-based model building workflows for drug discovery. This is particularly true for neglected tropical diseases since efforts to identify therapeutics for these diseases are often limited. In this report, we analyze the performance of several undersampling strategies in modeling the Dengue Virus 2 (DENV2) inhibitory activity, as well as the anti-flaviviral activities for the West Nile (WNV) and Zika (ZIKV) viruses. To this end, we build datasets comprising of 1218 (159 actives and 1059 inactives), 1044 (132 actives and 912 inactives) and 302 (75 actives and 227 inactives) molecules with known DENV2, WNV and ZIKV inhibitory activity profiles, respectively. We develop ensemble classifiers for these endpoints and compare the performance of the different undersampling algorithms on external sets. It is observed that data pruning algorithms yield superior performance relative to data selection algorithms. The best overall performance is provided by the one-sided selection algorithm with test set balanced accuracy (BACC) values of 0.84, 0.74 and 0.77 for the DENV2, WNV and ZIKV inhibitory activities, respectively. For the model building, we use the recently proposed GT-STAF information indices, and compare the predictivity of 3 molecular fragmentation approaches: connected subgraphs, substructure and alogp atom types, which are observed to show comparable performance. On the other hand, a combination of indices based on these fragmentation strategies enhances the predictivity of the built ensembles. The built models could be useful for screening new molecules with possible DENV, WNV and ZIKV inhibitory activities. ADMET modelers are encouraged to adopt undersampling algorithms in their workflows when dealing with imbalanced datasets.

Graphic abstract



Keywords Dengue virus · West nile virus · Zika virus · Undersampling · Support vector machine · Information index

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-019-00255-3) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Abbreviations

DENV	Dengue virus
WNV	West nile virus
ZIKV	Zika virus

Published online: 26 November 2019

GT-STAF IFI	Graph Theoretical Thermodynamic STAte Functions Information Index
QSAR	Quantitative structure–activity relationships

Introduction

Neglected tropical diseases (NTDs), also known as forgotten diseases, are communicable diseases that affect more than 1 billion people in developing countries of the tropical and sub-tropical regions [1]. Examples of NTDs include dengue, chikungunya, leishmaniasis, trypanosomiasis and chagas, among others. Typically, NTDs lack effective treatment and research efforts to identify therapeutics for these diseases are often limited.

In the specific case of dengue fever (DF), this is a viral mosquito-borne disease affecting over 390 million people worldwide. Many Dengue Virus (DENV) infections are asymptomatic, although cases that result in clinical manifestations may exasperate into potentially lethal forms of DF, denominated as dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), which are responsible for approximately 22,000 deaths annually [2].

Although genotypically related, there are five antigenically different DENV serotypes, denominated as DENV1, DENV2, DENV3, DENV4 and the recently discovered DENV5 [3]. Individuals infected by DENV gain immunity against the serotype responsible for the infection but there is limited cross-serotype immunity. On the contrary, an infection by a particular serotype seems to expose the patients to a more pronounced risk of developing severe forms of the disease on infection by a different serotype, a mechanism known as antibody-dependent enhancement (ADE) [4]. This clinical condition underscores the need for a broad-spectrum DENV therapy. Unfortunately, there is currently no clinically approved antiviral drug for DENV infections and thus palliative treatment is employed to overcome the symptoms of this disease. Recently, a tetravalent serotype DENV vaccine was approved for use in Brazil, Mexico and other countries in the endemic regions. However, there is already need to go back to the drawing board as a new serotype has appeared [5]. Additionally, there is clinical evidence suggesting that individuals to whom this vaccine has been administered may develop severe forms of DF upon their first DENV infection, which casts a cloud of uncertainty on its earnest applicability [6].

Efforts to develop small molecule or peptide therapeutics for DENV continue, although these have mainly used high throughput screening or followed in silico structure-based workflows with corresponding experimental validations. Nonetheless, these have not been translated into clinically useful therapeutics; for a detailed treatise of the recent academic and/or industrial undertaking towards the discovery of DENV therapeutics, see ref [7]. As it is often the case for neglected or new diseases, only few truly active DENV inhibitors have been reported in the literature, in a stark contrast to the 1000 s of inactive DENV inhibitors. Consequently, building robust and statistically meaningful DENV inhibitory activity models for virtual screening has been a challenge, which probably explains why ligand-based workflows for the discovery of DENV therapeutics are rare. Indeed, to the best of our knowledge, no DENV activity regression/classification model built over a structurally diverse dataset and considering all mechanisms of actions has been reported so far.

Imbalanced datasets, characteristically comprising of many more inactive compounds than the active ones, are a common challenge in ligand-based model building workflows. Often, a quick and simple solution to this challenge is balancing the chemical compound dataset by reducing the size of the set of inactive compounds, using techniques such as cluster analysis, principal component analysis or similarity/dissimilarity analysis [8]. However, these dimensionality reduction methods may result in an inefficient mapping of the inactive chemical compound space as the set of active compounds determines the cut-off to be employed. Cost-sensitive learning algorithms, such as penalized support vector machine (SVM) or Adacost (cost-sensitive AdaBoost ensemble), in which the misclassification of the active compounds is penalized have also been employed [9]. Nevertheless, since these follow a corrective feedback strategy, it is difficult to determine the influence of different class distributions on the classifier's performance [10, 11]. Other possible strategies for dealing with imbalanced datasets include: oversampling the active compounds or undersampling the inactive compounds. However, it has been argued that replicating active instances (oversampling) increases the risk of overfitting, while undersampling may result in an inadequate use of key chemical structural-activity information [12].

The goal of the present manuscript is to explore the performance of the different undersampling algorithms, i.e. random sampling, cluster centroids, near miss1, near miss2, near miss3, edited nearest neighbors, repeated edited nearest neighbors, all KNN and one sided selection, respectively, in modeling the DENV2 inhibitory activity. It should be noted that while undersampling algorithms have long existed in machine learning, their use in (Q)SPR (Quantitative Structure Property Relationships) modeling is rather uncommon. Indeed an extensive review of the literature yielded only a few reports in which undersampling was employed, and these were limited to the random undersampling algorithm exclusively [13–15].

Moreover, given the conserved nature of the DENV2 inhibition targets in pathogenic viruses of the Flaviviridae family, coupled with the recent resurgence and spread of

other flaviviral infections, particularly, Zika and West Nile fever in the Americas and Asia, we sought to also model the inhibitory activities for these flaviviruses, in order to compare the performance of the different undersampling algorithms considering different anti-flaviviral endpoints. Currently, no vaccine or effective antiviral treatment for WNV or ZKV infections is available for clinical use, although some molecular entities are currently under preclinical evaluation [16, 17]. It is thus essential that continued efforts to develop robust methods and tools for screening for new molecular entities with possible anti-flaviral activity are carried out.

One of the factors that may influence the performance of ligand-based models is the adequacy of the algorithms that characterize the chemical structural features. Herein, we employ the recently proposed GT-STAF (Graph Theoretical Thermodynamic STAte Functions) information indices, motivated by the metaphoric consideration of molecular structures as communication systems [18–22], to codify the chemical structural information. These information indices (IFIs) are based on the partition of molecular structures using graph-theoretic, molecular fingerprints and atomic physicochemical properties models, followed by the analysis of the statistical patterns of the obtained sets of molecular fragments. Although previous studies on the performance of these IFIs showed comparable to superior performance relative to popular academic and commercial descriptor computing software, these studies were based on small and congeneric datasets [18–24]. In this sense, this is the first study in which the utility of the proposed IFIs in codifying relevant chemical structural information is evaluated on a large and diverse chemical compound dataset.

Materials and methods

Chemical compound dataset

A detailed review of the literature was performed and chemical compounds with known DENV2, WNV and ZIKV activity profiles retrieved. In order to guarantee adequate dataset quality, the results obtained from raw natural product extracts were not considered due to the possible interference in the assays. Moreover, virtual screening results that lacked the corresponding experimental validation were discarded. The PubChem repository was also explored and chemical compounds with reported DENV2 and WNV activity profiles extracted from high throughput assays with clearly defined and validated endpoints; the chemical compound dataset for ZIKV inhibitory activity was retrieved from ref [25]. For the DENV2 and WNV inhibitory activities, compounds with IC_{50} or EC_{50} values ≤ 10 μM were considered as active, while those with IC_{50} or EC_{50} values > 10 μM were flagged as inactive. As for ZIKV a less stringent cutoff was employed, i.e. compounds with $AC_{50} \leq 20$ μM were considered as active, otherwise they were flagged as inactive.

Descriptor calculation and feature selection

The chemical compounds comprising the built datasets were characterized using a series of IFIs implemented in the freely available GT-STAF module of the ToMoCOMD-CARDD (acronym for Topological Molecular Computational Design-Computer Aided Rational Drug Design) software [26]. The following configurations were considered in the computation of the GT-STAF IFIs: three molecular fragmentation models, i.e. connected subgraphs (CS), substructure fingerprints (SS) and hydrophobicity atom-types (ALOGP), representative of graph-theoretic, molecular fingerprints and atomic physicochemical properties models, respectively, were selected. For these fragmentation models, unweighted and weighted, total and local, Shannon's entropy and mutual information indices were computed. Additionally, generalizations of the linear combination of atomic contributions to obtain global molecular indices were employed based on several aggregation operators stratified in four groups, i.e. norms, means, statistical invariants and classical algorithms. For a detailed treatise of the theoretical structure of the GT-STAF indices, see Ref. [27].

As it may be anticipated, numerous GT-STAF IFIs were obtained and thus dimensionality reduction procedures were necessary. Firstly, a Pearson's correlation coefficient-based filter, using a R^2 cutoff value of 0.60 was employed and thus retaining less correlated variables for each molecular fragmentation model. Subsequently, the information gain (IG), a supervised feature selection filter, was applied to the retained set of variables and ranked according to their discriminative capacity of the active compounds from the inactives. Finally, the best 100 variables for the CS, SS and ALOGP fragmentation models, as well as a combination of these (i.e. all GT-STAF indices) were selected for the subsequent ensemble model building. For the IG based feature selection, the IMMAN software was employed [28].

Ensemble model building and undersampling algorithms

In the present study, we employed the support vector machine (SVM) and the radial base function as learning algorithm and kernel method, respectively. Figure 1 is a scheme illustrating the workflow followed in the ensemble SVM model construction.

A tenfold random splitting of the built dataset into training (75%) and external validation (25%) sets was performed and for each fold, the training set was subject to an *n*-fold undersampling [n = imbalance ratio × 2; imbalance ratio = n(inactives)/n(actives)] using the following strategies: random

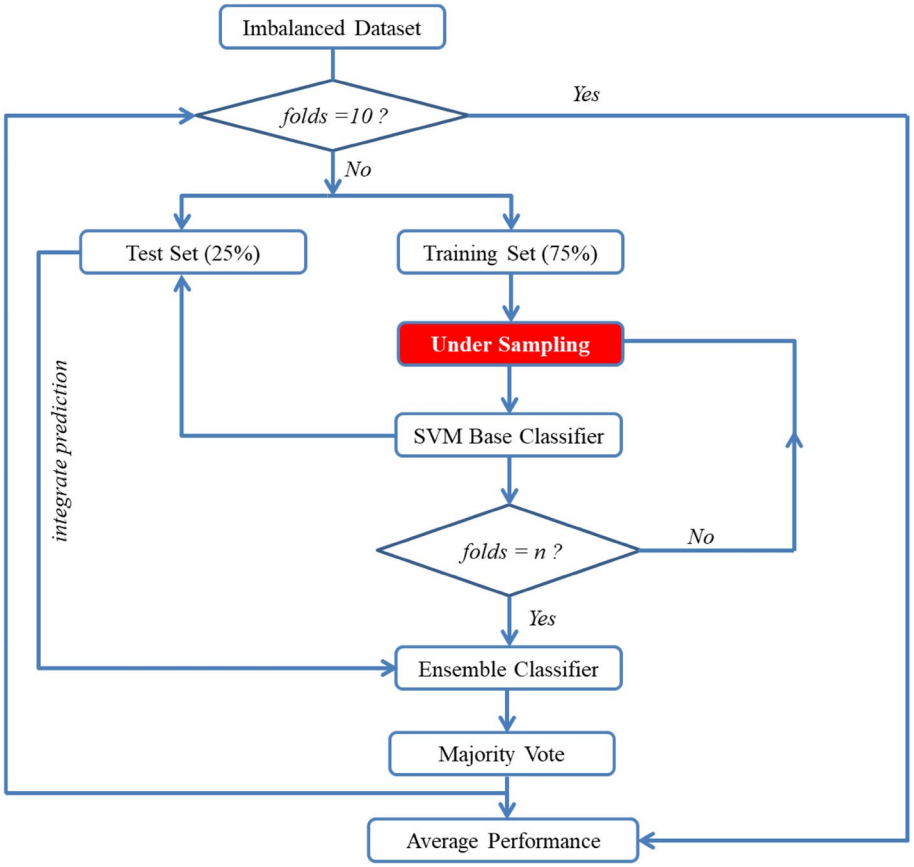


Fig. 1 Workflow of the undersampling protocol followed in the present study. This procedure was carried out for each of the nine undersampling strategies evaluated herein, based on the CS, SS and ALOGP molecular fragmentation models, as well as a combination of these; n = imbalance ratio × 2

sampling, cluster centroids, near miss1, near miss2, near miss3, edited nearest neighbors, repeated edited nearest neighbors, all KNN (k - nearest neighbors) and one sided selection. *Random undersampling* is a quick and straightforward approach for obtaining a balanced dataset in that a subset of instances is randomly selected for the targeted classes. On the other hand, the *cluster centroids* strategy employs the K-means algorithm to stratify the data into clusters, and subsequently use the centroids generated by this algorithm (rather than the original samples) to represent the obtained clusters with a reduced number of instances. The near miss strategies are based on the nearest neighbors algorithm [29]; for the *near miss1* variant a set of N instances in the larger class with the smallest average distance relative to the closest instances of the minority class is selected, while for the *near miss2* approach the smallest average distance to the farthest instances of the minority class is considered. As for the near miss3, given an instance in the minority class, a set of K nearest-neighbors is selected and then instances in the majority class are selected based on the largest average distance to the N nearest neighbors. The *edited nearest neighbors* and *repeated edited nearest neighbors* strategies employ the nearest-neighbors algorithm to eliminate instances from the larger class that are not consistent with a given neighborhood. If the nearest neighbors belong the same class then the instance is retained, otherwise it is removed from the dataset. In the case of the latter, iterations over this algorithm are performed, generally resulting in the removal of more instances from the dataset. The *all KNN* strategy differs from the *repeated edited nearest neighbors* in that the number of nearest neighbors to be evaluated is increased for each iteration [21]. In the *one sided selection* method, only examples in the majority class are removed while all minority class examples are kept. This method uses the 1-nearest neighbor rule and the Tomek links to prune out examples from the majority class. Tomek links are defined as follows: Given two instances, active and inactive, let δ (active, inactive) denote the distance between the 2 instances of different classes. The pair (active, inactive) is designated as a Tomek link, if there is no any instance (conveniently denominated *mol*), whose δ (active, *mol*) < δ (active, inactive) or δ (*mol*, inactive) < δ (inactive, active). In the one-sided selection method, the instances in the majority class involved in the Tomek links are pruned out since they could be noisy or borderline examples.

For each data subset, obtained following the aforementioned undersampling algorithms, an SVM base classifier was constructed, which in turn contributed to the ensemble SVM classifier. The majority vote criterion was selected as the consensus scheme.

Ensemble model validation

The constructed ensemble classifiers were subjected to a tenfold external validation using data sets obtained prior to the undersampling protocol (i.e. the validation sets were not obtained from the resulting under sampled data). In our opinion, this approach provides an earnest validation of the built models, since some undersampling strategies (e.g. the edited nearest neighbor and one sided selection algorithms) apply some sort of supervised stratification of the data. The performance of the ensemble classifiers was assessed using the classification metrics: balanced accuracy (BACC), Mathew's correlation coefficient (MCC) and Precision (PR), respectively. These classification metrics are defined as follows:

$$\begin{aligned} \text{BACC} &= \frac{1}{2}\left[\frac{TP}{(TP+FN)} + \frac{TN}{(TN+FP)}\right] \\ &= \frac{1}{2}\left[\text{Sensitivity} + \text{Specificity}\right] \end{aligned} \quad (1)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

$$PR = \frac{TP}{TP+FP} \quad (3)$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. The average BACC, MCC and PR values were computed as the arithmetic mean over the tenfold external validation sets. Moreover, the application domain (AD) of each classifier in the ensemble was determined using the leverage criterion, based on the Euclidean distance relative to the center of the molecular descriptors' space. For each model in the ensemble, the number of compounds within the respective AD was determined yielding a vector of values. Subsequently, the minimum value in the vector was defined as the number of compounds within the ensemble's AD. All the undersampling workflows, as well as the SVM model building and validation were performed using an in-house python script, which has been provided as Supplementary information S1.

Results

Chemical compound datasets

For the DENV inhibitory activity, a dataset comprising of 1218 compounds (159 actives and 1059 inactives), was retrieved from an extensive literature review [7, 30]. As it is evident, this dataset is characteristically imbalanced, with an imbalance ratio of 7 (1059/159). The set of active compounds comprises all mechanisms of DENV2 inhibitory activity reported to date: entry inhibitors, NS3 helicase inhibitors, NS2B-NS3 protease inhibitors, NS4B inhibitors, capsid protein inhibitors, NS5 methyltransferase inhibitors, NS5 Polymerase inhibitors, NS5 nuclear localization blockers, ligands targeting host cell factors and a subset of ligands without confirmed mechanisms.

As for the WNV, a dataset of 1044 compounds (132 actives and 912 inactives) was constructed, yielding an imbalance ratio of 7 (912/132) [31–36]. The active compounds included NS2B-NS3 proteinase inhibitors, NS5 N-terminal capping enzyme (CE) activity inhibitors and a series of inhibitors with undetermined mechanisms as they were identified using the Vero E6 cells viability assay.

The dataset for ZIKV inhibitory activity comprised of 302 compounds (75 actives and 227 inactives) screened for their caspase-3 inhibitory activity; this dataset had an imbalance ratio of 3 (227/75). The actives included compounds with neuroprotective activity (thus capable of preventing ZIKV induced cell death) and suppressors of ZIKV replication (i.e. antivirals).

As may be evident, these datasets comprise chemical compounds encompassing several activity mechanisms, which is a desirable attribute in that models built over a diverse bioactivity space favor the identification of new multi-target molecular entities through virtual screening experiments.

Feature selection and ensemble SVM classifiers

Ensemble SVM classifiers

For each inhibitory activity, balanced datasets were randomly extracted for each event (i.e. CS, SS, ALOP, respectively), as well as a combination of these. Subsequently, each balanced dataset (denominated One Sample henceforth) was employed to determine the optimum number of features for the SVM base classifiers, as well as the kernel configuration parameters (i.e. γ and C). These model settings were then fixed for the ensuing comparisons of the different undersampling strategies. Table 1 shows the number of base classifiers comprising the ensemble models for each event and their combinations (all GT-STAF indices), as well as the corresponding γ and C parameters for the SVM base classifiers.

Table 1 Configuration parameters and number of SVM base classifiers comprising the ensemble models for the DENV, WNV and ZIKV inhibitory activity

	No of base classifiers	Gamma (γ)	C
DENV inhibitory activity			
CS	14	0.039	6.0
SS	14	0.051	6.0
ALOGP	14	0.050	4.0
All GT-STAF	14	0.039	8.0
WNV inhibitory activity			
CS	14	0.076	6.0
SS	14	0.061	4.0
ALOGP	14	0.056	5.0
All GT-STAF	14	0.056	5.0
ZIKV inhibitory activity			
CS	6	0.024	8.0
SS	6	0.027	8.0
ALOGP	6	0.036	7.0
All GT-STAF	6	0.024	8.0

DENV inhibitory activity

Table 2 shows the average performance of the DENV inhibitory activity ensemble models on external validation sets based on the BACC, MCC and PR, respectively, for the three molecular fragmentation models, as well as combinations of these.

It is interesting to note that while generally comparable BACC (0.73–0.84) is obtained for all SVM models, dissimilar performance is observed for the other validation parameters, underscoring the role of the different undersampling strategies in the performance of the models. In the case of the CS-based models, the random, cluster centroids and near miss undersampling strategies yield good BACC (0.75–0.85) and modest MCC (0.35–0.55), but their PR values (0.28–0.47) are below the limit of acceptability, an indication of a high number of false positives (see Eq. 3). On the other hand, the edited nearest neighbors, repeated edited nearest neighbors, all KNN and one sided selection yield good performance for all metrics (ACC = 0.80–0.83, MCC = 0.65–0.70 and PR = 0.72–0.88 SP = 0.99), with a slightly better performance provided by the One Sided Selection algorithm (ACC = 0.80, MCC = 0.70, PR = 0.88). Overall, the near miss approaches yield the least favorable performance for all the considered parameters.

Likewise, the SS molecular fragmentation models based on the random and cluster centroid undersampling approaches yield good BACC (0.82–0.83) and MCC (0.67–0.78) values, while their corresponding PR values (0.44–0.45) are below the random threshold. For the near miss algorithms, with the exception of the good BACC values (0.76–0.77), modest to rather poor performance is observed with the rest of the parameters, with MCC = 0.34–0.38 and PR = 0.26–0.28, respectively. Conversely, the edited nearest neighbors, repeated edited nearest neighbors, all KNN and one sided selection yield good quality classifiers (BACC = 0.80–0.81, MCC = 0.64–0.70 and PR = 0.66–0.87), with the best performance provided by the one sided selection algorithm (BACC = 0.80, MCC = 0.70, PR = 0.87). In the case of the ALOGP model, the random, cluster centroid and near miss algorithms yield good BACC (0.73–82) and modest MCC (0.32 - 0.49), while the PR is once again below the limit of acceptability (0.26–0.40). As in the CS and SS models, the edited nearest neighbors, repeated edited nearest neighbors, all KNN and one-sided selection yield the best performance with BACC = 0.77–0.82, MCC = 0.57–0.65 and PR = 0.55–0.86. Similar to the aforementioned models, the classifiers comprising all the GT-STAF index types show modest to good BACC (0.78–0.85) and MCC (0.39–0.57) for the random, cluster centroid and near miss algorithms, while the PR values (0.30–0.49) are unsatisfactory. On the other hand, the edited nearest neighbors, repeated edited nearest neighbors, all KNN and one sided selection yield robust ensembles with BACC = 0.84–0.86, MCC = 0.71–0.72 and PR = 0.74–0.83, respectively. Note that the best overall DENV inhibitory activity classifier is provided by the one sided selection algorithm based on all GT-STAF indices with BACC = 0.84, MCC = 0.72 and PR = 0.83, respectively.

WNV inhibitory activity

Table 3 shows the average performance of the WNV inhibitory activity ensemble models assessed over tenfold external validation sets in terms of the BACC, MCC and PR, respectively.

For the CS-based classification models, it is observed that while the random and cluster centroids algorithms yield acceptable BACC (0.73–0.77) and MCC (0.34–0.42), their PR values (0.31–0.37) are low, indicating a higher number of FP (see Eq. 3). The near miss algorithms-based ensembles are of even much lower quality, evidenced by the rather unfavorable PR values (0.14–0.15) and nearly random classification performance (MCC = 0.16–0.17). An improvement of the MCC and PR is achieved with the nearest neighbor algorithms, although only the edited nearest neighbors algorithm achieves globally acceptable classifiers (BACC = 0.73, MCC = 0.48, PR = 0.57). The best overall performance for the CS-based classifiers is obtained with the one-sided selection algorithm with BACC = 0.73, MCC = 0.53 and PR = 0.69.

For the SS-based ensemble models, only the edited nearest neighbors and one-sided selection provide PR values above the limit of acceptability, i.e. 0.56 and 0.75,

Table 2 Test set validation parameters of the built ensemble classifiers for predicting the DENV2 inhibitory activity, based on nine undersampling algorithms and the CS, SS and ALOGP molecular fragmentation models, as well as a combination of these (all GT-STAF indices)

	Balanced Acc.	MCC	Precision	AD[a]
Connected sub-graphs				
One sample	0.83	0.50	0.41	291
Random	0.85	0.55	0.47	294
Cluster centroids	0.81	0.46	0.46	299
Near Miss1	0.80	0.43	0.33	288
Near Miss2	0.78	0.39	0.31	288
Near Miss3	0.75	0.35	0.28	289
Edited nearest neighbors	0.80	0.65	0.75	300
Repeated edited nearest neighbors	0.83	0.67	0.72	299
All KNN	0.81	0.66	0.75	300
One sided selection	0.80	0.70	0.88	298
Substructure fingerprints				
One sample	0.80	0.47	0.40	295
Random	0.82	0.51	0.45	295
Cluster centroids	0.83	0.51	0.44	297
Near Miss1	0.77	0.38	0.29	293
Near Miss2	0.76	0.36	0.28	294
Near Miss3	0.76	0.34	0.26	293
Edited nearest neighbors	0.81	0.67	0.77	294
Repeated edited nearest neighbors	0.83	0.64	0.66	291
All KNN	0.83	0.65	0.69	293
One sided selection	0.80	0.70	0.87	293
Alogp atom types				
One sample	0.81	0.47	0.38	295
Random	0.82	0.49	0.40	295
Cluster centroids	0.81	0.45	0.37	298
Near Miss1	0.75	0.34	0.28	293
Near Miss2	0.75	0.36	0.30	291
Near Miss3	0.73	0.32	0.26	290
Edited nearest neighbors	0.77	0.57	0.67	299
Repeated edited nearest neighbors	0.82	0.58	0.55	298
All KNN	0.79	0.56	0.60	299
One sided selection	0.77	0.65	0.86	298
All GT-STAF indices				
One sample	0.82	0.49	0.41	295
Random	0.85	0.57	0.49	295
Cluster centroids	0.85	0.55	0.45	298
Near Miss1	0.80	0.43	0.35	293
Near Miss2	0.78	0.40	0.32	292
Near Miss3	0.78	0.39	0.30	292
Edited nearest neighbors	0.86	0.71	0.74	294
Repeated edited nearest neighbors	0.84	0.68	0.73	296
All KNN	0.84	0.68	0.72	296
One sided selection	**0.84**	**0.72**	**0.83**	**296**

Parameters of the best models are indicated in bold

[a]*AD* applicability domain, out of 305 test set compounds

respectively. The rest of the parameters are similar to those obtained by the CS-based models, with the near miss algorithms providing the worst overall performance.

A similar trend is observed with the ALOGP and all GT-STAF indices ensemble models with only the edited nearest neighbors, all KNN and one-sided selection yielding

Table 3 Test set validation parameters of the built ensemble classifiers for predicting the WNV inhibitory activity, based on nine undersampling algorithms and the CS, SS and ALOGP molecular fragmentation models, as well as a combination of these (all GT-STAF indices)

	Balanced Acc.	MCC	Precision	AD[a]
Connected sub-graphs				
One sample	0.75	0.36	0.30	248
Random	0.77	0.42	0.37	248
Cluster centroids	0.73	0.34	0.31	251
Near Miss1	0.61	0.17	0.17	195
Near Miss2	0.61	0.16	0.15	196
Near Miss3	0.62	0.17	0.15	190
Edited nearest neighbors	0.73	0.48	0.59	255
Repeated edited nearest neighbors	0.73	0.43	0.47	255
All KNN	0.73	0.42	0.46	255
One sided selection	0.73	0.53	0.69	255
Substructure fingerprints				
One sample	0.71	0.29	0.27	249
Random	0.72	0.32	0.29	249
Cluster centroids	0.73	0.33	0.29	254
Near Miss1	0.59	0.13	0.16	211
Near Miss2	0.60	0.14	0.15	209
Near Miss3	0.62	0.17	0.16	217
Edited nearest neighbors	0.74	0.48	0.56	257
Repeated edited nearest neighbors	0.72	0.36	0.38	256
All KNN	0.72	0.41	0.45	256
One sided selection	0.72	0.51	0.68	257
Alogp atom types				
One sample	0.71	0.31	0.29	242
Random	0.74	0.37	0.35	242
Cluster centroids	0.75	0.34	0.28	253
Near Miss1	0.62	0.16	0.16	207
Near Miss2	0.62	0.16	0.17	212
Near Miss3	0.61	0.15	0.17	208
Edited nearest neighbors	0.76	0.52	0.58	252
Repeated edited nearest neighbors	0.75	0.42	0.41	252
All KNN	0.75	0.49	0.55	250
One sided selection	**0.74**	**0.58**	**0.76**	**251**
All GT-STAF Indices				
One sample	0.74	0.35	0.31	251
Random	0.76	0.40	0.34	251
Cluster centroids	0.73	0.34	0.31	254
Near Miss1	0.61	0.16	0.18	211
Near Miss2	0.59	0.13	0.15	205
Near Miss3	0.61	0.16	0.16	207
Edited nearest neighbors	0.75	0.50	0.57	256
Repeated edited nearest neighbors	0.75	0.46	0.48	257
All KNN	0.76	0.48	0.50	256
One sided selection	0.74	0.53	0.64	256

Parameters of the best models are indicated in bold

[a]*AD* applicability domain, out of 261 test set compounds

PR parameters superior to random performance (> 50%). The PR values for the ALOGP-based classifiers are 0.58 (edited nearest neighbors), 55 (all KNN) and 0.76 (one-sided selection), while those for the GT-STAF-based ensembles are 0.58 (edited nearest neighbors), 0.50 (all KNN) and 0.64 (one-sided selection), respectively.

ZIKV inhibitory activity

Table 4 illustrates the average performance of the ZIKV inhibitory activity ensemble models evaluated over tenfold external validation sets and expressed in terms of the classification parameters BACC, MCC and PR, respectively.

The CS-based ensembles for ZIKV inhibitory activity yields satisfactory performance for the following algorithms:

Table 4 Test set validation parameters of the built ensemble classifiers for predicting the ZIKV inhibitory activity, based on nine undersampling algorithms and the CS, SS and ALOGP molecular fragmentation models, as well as a combination of these (all GT-STAF indices)

	Balanced Acc.	MCC	Precision	AD[a]
Connected sub-graphs				
One sample	0.72	0.40	0.48	71
Random	0.76	0.47	0.53	71
Cluster centroids	0.79	0.49	0.47	73
Near Miss1	0.68	0.32	0.43	63
Near Miss2	0.66	0.28	0.42	61
Near Miss3	0.65	0.26	0.37	61
Edited nearest neighbors	0.75	0.45	0.51	74
Repeated edited nearest neighbors	0.67	0.30	0.34	72
All KNN	0.72	0.38	0.41	73
One sided selection	0.72	0.45	0.57	73
Substructure fingerprints				
One sample	0.68	0.32	0.41	72
Random	0.69	0.34	0.44	72
Cluster centroids	0.72	0.39	0.45	72
Near Miss1	0.66	0.28	0.41	64
Near Miss2	0.64	0.25	0.38	62
Near Miss3	0.60	0.17	0.31	64
Edited nearest neighbors	0.67	0.27	0.34	73
Repeated edited nearest neighbors	0.66	0.31	0.36	73
All KNN	0.66	0.28	0.35	72
One sided selection	0.72	0.46	0.65	73
Alogp atom types				
One sample	0.70	0.35	0.41	70
Random	0.73	0.41	0.45	70
Cluster centroids	0.716	0.38	0.45	72
Near Miss1	0.63	0.22	0.35	67
Near Miss2	0.63	0.23	0.35	66
Near Miss3	0.65	0.25	0.34	66
Edited nearest neighbors	0.73	0.41	0.43	73
Repeated edited nearest neighbors	0.68	0.33	0.35	73
All KNN	0.71	0.36	0.37	73
One sided selection	0.69	0.38	0.52	74
All GT-STAF indices				
One sample	0.76	0.45	0.49	67
Random	0.75	0.44	0.51	67
Cluster centroids	0.73	0.41	0.48	69
Near Miss1	0.68	0.32	0.43	58
Near Miss2	0.68	0.32	0.42	58
Near Miss3	0.66	0.29	0.41	56
Edited nearest neighbors	0.73	0.39	0.42	72
Repeated edited nearest neighbors	0.68	0.32	0.38	71
All KNN	0.71	0.37	0.43	72
One sided Selection	**0.77**	**0.53**	**0.66**	**73**

Parameters of the best models are indicated in bold

[a]*AD* applicability domain, out of 76 test set compounds

random (BACC = 0.76, MCC = 0.47, PR = 0.53), edited nearest neighbors (BACC = 0.75, MCC = 0.45, PR = 0.51) and one-sided selection (BACC = 0.72, MCC = 0.45, PR = 0.57). The models based on the rest of the undersampling algorithms are inconsistent with the limits of model acceptance.

For the SS and ALOGP molecular fragmentation methods, only the ensemble based on the one-sided selection algorithm provides parameters above the limit of acceptability, *i.e.* BACC = 0.72, MCC = 0.46 and PR = 0.65 in the case of the former, and BACC = 0.69, MCC = 0.38 and PR = 0.52 for the latter. Lastly, the models yield based on a combination of GT-STAF indices yield good predictivity for the random (BACC = 0.75, MCC = 0.44 and PR = 0.51) and one sided selection (BACC = 0.77, MCC = 0.53 and PR = 0.66). Note that for all the models (based on the three molecular fragmentation approaches and combinations of these), the one-sided selection algorithm yields the best overall undersampling algorithm.

Applicability domain

Tables 2, 3, 4 illustrates the number of compounds within the AD for each of the reported ensemble classifiers. For the DENV inhibitory activity, the built models flagged 94–98% of the test set compounds as within the corresponding ADs. In the case of the WNV inhibitory activity classifiers, these have more dispersed AD ranges with 73–98% of the test compounds found to lie within the respective ADs; the near miss algorithms based on the CS molecular fragmentation approach yield the lowest percentages (73–75%) of test compounds within the models' ADs. Finally, the ZIKV inhibitory activity classifiers labelled 74–97% of the test set compounds as lying within the respective ADs, with the near miss algorithms based on all GT-STAF indices associated with the lowest percentages (74–76%). Overall, the built models encompassed the majority of the test compounds within the corresponding ADs. It is also, important to note that besides yielding poor quality classifiers, the near miss algorithms had the highest number of test compounds (approx. 25%) identified to lie outside the respective ADs.

Discussion

Undersampling algorithms

From the obtained results, a distinction in performance is observed between data selection (i.e. random sampling, cluster centroids and near miss) and the data pruning (i.e. edited nearest neighbors, all KNN and one sided selection) algorithms for the modeled properties, with generally superior performance for the latter. The high number of false positives observed with the data selection algorithms is attributed to numerous boundary examples, which tend to induce an incorrect adjustment of the decision hyperplane, resulting in false predictions. Indeed, the near miss algorithms which show the worst performance for the three properties modeled herein are in general designed to favor the selection of sets of instances with the smallest average distances to the minority class which consequently produces numerous boundary examples for both the minority and majority classes. In the case of the cluster centroid algorithm, boundary examples are represented by similar "artificial examples" (centroids), and thus the same challenge persists. Moreover, it is important to highlight that the very use of centroids could result in loss of chemical structural formation. In fact, it is not clear what these artificial examples could mean from a molecular topostructural or topochemical perspective. As regards the random undersampling, the poor performance could be attributed to both the removal of potentially relevant data for model building and the presence of borderline examples.

On the other hand, the data pruning methods follow a supervised approach based on the nearest neighbor algorithm in which wrongly classified examples (typically borderline or noisy examples) in the larger class are flagged for removal. As a result, these algorithms yield non-overlapping distributions, which are suitable for classification model building. The findings reported herein are consistent with some previous reports in the literature [37, 38].

Molecular fragmentation models

Finally, in the analysis of the different molecular fragmentation models, comparable performance is obtained with the CS, SS and ALOGP approaches, notwithstanding the minor edge for the CS approach in the case of the DENV and ZIKA inhibitory activities, and ALOGP for the WNV inhibitory activity ensembles. Moreover, combining the different molecular fragmentation approaches (i.e. all GT-STAF indices) enhances the performance of these indices in modeling the considered flaviviral inhibitory activities. Based on the satisfactory results reported here in, it may be inferred that GT-STAF IFIs codify important chemical structural information useful in the modeling of the bioactivity of chemical compounds.

Conclusions

Machine learning methods have increasingly gained greater acceptability in drug discovery and molecular modeling due to their capacity to provide accurate predictions

of the ADMET profiles of chemical compounds. However, machine learning-based models are, among other factors, "as good as the data used to build them". This becomes an enormous challenge when dealing with the so-called neglected diseases, since often few positive training examples are available. Moreover, with the usually high attrition rates in drug discovery initiatives [39], negative training examples are often several folds more than the positive ones. Consequently, modelers or drug hunters regularly need to deal with imbalanced datasets. While the option usually adopted is to discard majority class examples through random or dimensionality reduction procedures to obtain a balanced dataset, this results in the loss of volumes of potentially valuable information on profiles of negative ligands. In the present manuscript, we compare the performance of several undersampling criteria in the modeling of the DENV2, WNV and ZIKV inhibitory activities. It is found that the data pruning undersampling algorithms (edited nearest neighbors, all KNN and one sided selection) generally provide superior performance if compared to that obtained with the data selection algorithms (random sampling, cluster centroids and near miss), with the best overall performance provided by the one-sided selection algorithm, while the near miss algorithms yield the worst performance for all the three bioactivities. The superior performance of the data pruning algorithms is attributed to their inherent capacity to eliminate borderline and/or noisy examples along the decision boundary, while the near miss algorithms in principle favor the selection of examples at closer distances to the minority class. Moreover, a comparison of the CS, SS and ALOGP molecular fragmentation models is carried out and similar statistical quality is observed, while combining the three approaches enhances the performance of the GT-STAF formalism. In light of the good performance obtained with the ensemble classifiers, it is our opinion that the results here obtained should stimulate screening initiatives for new molecular entities with possible DENV2, WNV and ZIKV inhibitory activities. Finally, drug hunters and ADMET modelers, in general, are encouraged to adopt undersampling algorithms in their modeling workflows.

Acknowledgements The authors appreciate the reviewers for their valuable comments and taking the time to revise the submitted python scripts.

References

1. Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, Savioli L (2007) N Engl J Med 357(10):1018
2. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O (2013) Nature 496(7446):504
3. Normile D (2013) Science 342(6157):415
4. Guzman MG, Alvarez M, Halstead SB (2013) Arch Virol 158(7):1445
5. Capeding MR, Tran NH, Hadinegoro SRS, Ismail HIHM, Chotpitayasunondh T, Chua MN, Luong CQ, Rusmil K, Wirawan DN, Nallusamy R (2014) Lancet 384(9951):1358
6. Normile D (2017) Science 358:1514
7. Behnam MA, Nitsche C, Boldescu V, Klein CD (2016) J Med Chem 59(12):5622
8. Brito-Sánchez Y, Marrero-Ponce Y, Barigye SJ, Yaber-Goenaga I, Morell Perez C, Le-Thi-Thu H, Cherkasov A (2015) Mol Inform 34(5):308
9. Barigye SJ, Freitas MP, Ausina P, Zancan P, Sola-Penna M, Castillo-Garit JA (2018) ACS Comb Sci 20(2):75
10. Hoens TR, Chawla NV (2013) Imbalanced datasets: from sampling to classifiers. In: Haibo H, Yunqian M (eds) Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press, New Jersey, p 43
11. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer, Berlin
12. He G, Han H, Wang W (2005) An over-sampling expert system for learing from imbalanced data sets. 2005 International Conference on Neural Networks and Brain: IEEE, p 537
13. Newby D, Freitas AA, Ghafourian T (2013) J Chem Inform Model 53(2):461
14. Gadaleta D, Manganelli S, Roncaglioni A, Toma C, Benfenati E, Mombelli E (2018) J Chem Inform Model 58(8):1501
15. Zang Q, Rotroff DM, Judson RS (2013) J Chem Inform Model 53(12):3244
16. Morens DM, Fauci AS (2017) J Infect Dis 216(suppl_10):S857
17. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A (2017) Nature 546(7658):411
18. Barigye SJ, Marrero-Ponce Y, Martínez-López Y, Martínez-Santiago O, Torrens F, García-Domenech R, Galvez J (2012) SAR QSAR Environ Res 24(1):3–34
19. Barigye SJ, Marrero-Ponce Y, Alfonso-Reguera V, Pérez-Giménez F (2013) Chem Phys Lett 570:147
20. Barigye SJ, Marrero-Ponce Y, Martínez-López Y, Torrens F, Artiles-Martínez LM, Pino-Urias RW, Martínez-Santiago O (2013) J Comp Chem 34(4):259
21. Barigye SJ, Marrero-Ponce Y, Martínez-Santiago O, Martínez-López Y, Torrens F (2013) Curr Comput Aided Drug Des 9:164
22. Barigye SJ, Marrero-Ponce Y, Pérez-Giménez F, Bonchev D (2014) Mol Divers 18(3):673
23. Barigye SJ, Marrero-Ponce Y, Zupan J, Pérez-Giménez F, Freitas MP (2014) Bull Chem Soc Jpn 88(1):97
24. Marrero-Ponce Y, Santiago OM, López YM, Barigye SJ, Torrens F (2012) J Comput Aided Mol Des 26(11):1229
25. Xu M, Lee EM, Wen Z, Cheng Y, Huang W-K, Qian X, Julia T, Kouznetsova J, Ogden SC, Hammack C (2016) Nat Med 22(10):1101
26. He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. Wiley, Hoboken
27. Barigye SJ, Marrero-Ponce Y (2016) Digital communication and chemical structure codification. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, Berlin, p 1
28. Urias RWP, Barigye SJ, Marrero-Ponce Y, García-Jacas CR, Valdes-Martiní JR, Perez-Gimenez F (2015) Mol Divers 19:305
29. Mani I, Zhang I (2003) kNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets
30. National Center for Biotechnology Information. Southern Research Specialized Biocontainment Screening Center. PubChem Database https://pubchem.ncbi.nlm.nih.gov/bioassay/540333. Accessed 17 Apr 2019.

31. Goodell JR, Puig-Basagoiti F, Forshey BM, Shi P-Y, Ferguson DM (2006) J Med Chem 49(6):2127
32. Behnam MAM, Graf D, Bartenschlager R, Zlotos DP, Klein CD (2015) J Med Chem 58(23):9354
33. Aravapalli S, Lai H, Teramoto T, Alliston KR, Lushington GH, Ferguson EL, Padmanabhan R, Groutas WC (2012) Bioorg Med Chem 20(13):4140
34. National Center for Biotechnology Information. Southern Research Specialized Biocontainment Screening Center. PubChem Database https://pubchem.ncbi.nlm.nih.gov/bioassay/1650. Accessed 18 Apr 2019
35. National Center for Biotechnology Information. Southern Research Specialized Biocontainment Screening Center. PubChem Database https://pubchem.ncbi.nlm.nih.gov/bioassay/588371. Accessed 15 Apr 2019
36. National Center for Biotechnology Information. PubChem Database https://pubchem.ncbi.nlm.nih.gov/bioassay/1079778. Accessed 20 Apr 2019
37. Tomek I (1976) IEEE Trans Syst Man Cybern 6(6):448
38. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. Nashville, Icml, p 179
39. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, Pairaudeau G, Pennie WD, Pickett SD, Wang J (2015) Nat Rev Drug Discov 14(7):475

Affiliations

Stephen J. Barigye[1] · José Manuel García de la Vega[1] · Juan A. Castillo-Garit[2]

✉ Stephen J. Barigye
sjbarigye@gmail.com

[1] Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain

[2] Unidad de Toxicología Experimental, Universidad de Ciencias Médicas de Villa Clara, 50200 Santa Clara, Villa Clara, Cuba