molecular informatics
models – molecules – systems

Accepted Article

Title: Generative Adversarial Networks (GANs) Based Synthetic Sampling for Predictive Modeling

Authors: Stephen Jones Barigye, José Manuel García-de la Vega, and Yunierkis Perez-Castillo

To be cited as: *Mol. Inf.* 10.1002/minf.202000086

Link to VoR: https://doi.org/10.1002/minf.202000086

10.1002/minf.202000086

Generative Adversarial Networks (GANs) Based Synthetic Sampling for Predictive Modeling

Stephen J. Barigye,[†*] José Manuel García-de la Vega[†], Yunierkis Perez-Castillo,[§]

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

[§]Bio-Chemoinformatics Research Group and Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Quito, 170504, Ecuador

*corresponding author: sjbarigye@gmail.com

Abstract

In the present report we evaluate the possible utility of the Generative Adversarial Networks (GANs) in mapping the chemical structural space for molecular property profiles, with the goal of subsequently yielding synthetic (artificial) samples for ligand-based molecular modeling. Two case studies are considered: BACE-1 (β-Secretase 1) and DENV (Dengue Virus) inhibitory activities, with the former focused on data populating and the latter on data balancing tasks. We train GANs using subsamples extracted from datasets for each bioactivity endpoint, and apply the trained networks in generating synthetic examples from the respective bioactivity chemical spaces. Original and synthetic samples are pooled together and employed to build BACE-1 and DENV inhibitory activity classifiers and their performance evaluated over tenfold external validation sets. In both case studies, the obtained classifiers demonstrate satisfactory predictivity with the former yielding accuracy (ACC) and Mathew's correlation coefficient (MCC) values of 0.80 and 0.59, while the latter produces balanced accuracy(BACC) and MCC values of 0.81 and 0.70, respectively. Moreover, the statistics of these classifiers are compared with those of other models in the literature demonstrating comparable to better performance. These results suggest that GANs may be useful in mapping the chemical space for molecular property profiles of interest, and thus allow for the extraction of synthetic examples for computational modeling.

Key words: Generative Adversarial Network, β-Secretase 1, Dengue Virus, Machine Learning, Data Sparsity

Introduction

Machine learning methods have been employed for decades in the computational modeling of molecular properties and their interaction with biomolecules, particularly due to their ability to map complex relationships between chemical moieties and observed chemical and/or biological outcomes.[1] Indeed, the use of predictive data-driven workflows has reshaped research paradigms in chemistry, genomics, drug discovery, toxicological risk assessment and among others, favored in part by the volumes of data accessible in public repositories, as well as the advancement in computational power.

Until recently the most popular machine learning algorithms employed have been kernel-based methods (e.g. support vector machines, gaussian process) and tree-based ensemble algorithms (e.g. random forest, boosting trees), while neural networks (NN) seemed to be relegated to a "long winter" although they continued to be employed in chemoinformatics applications.[2] However, progress in supervised and unsupervised learning algorithms, coupled with state-of-the art high-end computing power, has allowed for the development deep neural networks (DNN) with remarkably improved performance relative to classical ML algorithms.[3] Indeed DNN have found applications in a wide array of fields including image analysis, speech recognition, fraud detection and virtual screening of potentially therapeutic molecular entities.[4] Examples of DNN include: convolutional neural networks (CNN) known to be useful in processing multiarray data such as images, recurrent neural networks (RNN) which have found utility particularly in speech recognition systems and machine translation, autoencoders (AE) (useful for dimensionality reduction) and generative adversarial networks (GAN), suitable for data generation.

The CNN, RNN and AE have found applications in several molecular modeling tasks such as predictive computational modeling, *de novo* design of chemical structures, analysis of ligand-macromolecule interactions, among others. Likewise GANs have been applied in the *de novo* design of protein structures and to generate small molecules based on binding pocket shape complementarity or using SMILES molecular representations as inputs.[5] However, to the best of our knowledge, GANs have not been extrapolated to (over)sampling paradigms as possible solutions to data scarcity, sparsity or imbalance in chemoinformatics, although applications to aimed at generating multiarray sample data such as medical and hyperspectral images for robust model training may be cited.[6] Clearly that there exists natural extrapolations to similar challenges in chemoinformatics paradigms where additional data samples are often required.

In the present report, we sought to assess the possible utility of GAN in computational molecular modeling, particularly in generating artificial data representative of a given chemical space without explicit knowledge of the corresponding molecular structures. Two case studies were considered herein: BACE-1 (β-Secretase 1) and DENV (Dengue Virus Inhibitory Activity) inhibitory activities, with the former focused on data (re)populating and the latter on data balancing tasks, respectively. One of the key strengths of the GANs approach for data generation is that it does not require the explicit modeling of the probability density function corresponding to the data space of interest. Other approaches followed to generate synthetic data include the SMOTE (Synthetic Minority Over-sampling Technique) and adaptive synthetic (ADASYN) sampling algorithms, which use the *k* nearest neighbors (KNN) algorithm. Nonetheless, as previously reported in the literature these KNN-based interpolation algorithms may present several weakness, e.g. are unsuitable for decision boundary samples, may generate duplicate samples in dense clusters or be biased towards noisy samples,[7] underscoring the need to

develop or apply alternative approaches for robust synthetic data generation. It should be noted that the aim of the present study is not to generate new chemical structures, but rather employ the GANs as a strategy to generate synthetic samples to (re)populate data matrices computed for given chemical compound datasets as a possible solution to data sparsity or imbalance.

Materials and Methods

Chemical Compound Datasets and Molecular Descriptor Calculation

A BACE-1 inhibitory activity dataset with the corresponding set of precomputed molecular descriptors was retrieved from the literature.[8] This dataset comprised of 1522 compounds with 825 inactives and 697 actives, respectively; the same binary cutoff adopted in ref [8] was employed (*i.e.* $IC_{50} \leq 100$ nM for actives). For this dataset, we sought to determine if samples generated for a given bioactivity chemical space using the GANs may be useful for robust model building.

On the other hand, with the goal of evaluating the possible utility of GANs in modeling tasks involving imbalanced data samples, a set of 1218 chemical compounds with known DENV inhibitory activity profiles (*i.e.* 159 actives and 1059 inactives, respectively) was obtained from a from a previous study.[9] For this dataset, we computed GT-STAF (acronym for Graph Theoretical Thermodynamic STAte Functions) information indices based on the substructure molecular fragmentation criteria. These indices, based on the analysis of chemical structures as communication systems, have been reported to provide satisfactory correlations with dissimilar molecular properties and codify orthogonal chemical structural information when compared to descriptors implemented in popular descriptor computing software.[9-10]

Generative Adversarial Networks , Datasets and Descriptor Computation

The GAN framework is comprised of two multilayer perceptron models: a generator (G) which produces samples and a discriminator (D) for distinguishing training data samples from those generated by G (model data distribution). The G is trained to maximize the probability of D misclassifying a generated sample as a real one, while D attempts to learn to distinguish generated samples form real ones, in analogy to a zero-sum game. The value function V(G, D) for the GAN is formulated as:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_{data}}}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_{model}}}[\log(1 - D(\boldsymbol{x}))] \qquad (\mathbf{1})$$

where p_{data} and p_{model} are the probability distributions for the data and generated samples, respectively. It follows that with sufficient capacity and learning time for G and D, an optimal global solution is achieved for which p_{data} converges to p_{model}.

The BACE-1 inhibitory activity data matrix was randomly split into 2 sets, with each set comprising 50% (761 compounds) of the original set. The first set was employed to train a GAN which was consequently used to generate the same number of artificial samples, while the second set was reserved for external validation. For the DENV inhibitory activity, only the sub-data matrix corresponding to the active chemical compounds (159) was extracted to train a GAN and subsequently used to generate artificial active samples to yield a balanced dataset.

The following architecture was adopted for G: two hidden layers, leaky ReLU (rectified linear unit) function for first/hidden layer non-linearity and tanh for the output layer function (output is an $n \times m$ data matrix, where n and m are the number of samples and variables, respectively). A similar configuration was employed for D, although with different dimensions (e.g. the output is a scalar, *i.e.* the equivocation probability) and output layer function (sigmoid), for details see supporting information for code. For the GAN training, Adam and binary cross-

entropy were employed as the model's stochastic optimization algorithm and loss function, respectively, while the learning rate was set to 0.0002. The number of training epochs was set to 50 and 10 for BACE-1 and DENV inhibitory activities, respectively.

Model Building and Validation

For the BACE-1 inhibitory activity, the training set (50% of original data matrix) and the GAN samples were assembled and used for the model building. The BACE+GAN dataset was split into training (75%) and test (25%) sets following a 10 fold validation procedure. Using the training set, the mutual information (MI) parameter provided by the sklearn module (Python scripting language) was employed to select the set of variables (45%) that best explain the BACE-1 inhibitory activity classification variable. Next, we explored a series of classification algorithms, using the genetic algorithm (GA) as the optimization strategy. The following configuration was employed: the F1 score was set as the objective function, the population size (number of solutions) was set to 100, the crossover and mutation ratios were 0.5 and 0.2, respectively. The best model, selected based on the F1 score, was then externally validated using the other 50% of original data matrix. Concurrently, the training set without the GAN samples was employed to build a classification model using the same set of variables and configuration for the best model and its performance evaluated over the external validation set.

As for the DENV inhibitory activity, the GAN balanced set was divided into training (75%) and external validation sets (25%), respectively. Similar to the BACE-1 inhibitory activity modeling, a MI filter was applied to the training set to extract the most relevant descriptors, followed by the GA-based model building. The predictivity of the best classification model was assessed following a 10-fold external validation procedure.

The following validation metrics were employed: accuracy (ACC), specificity (SP), sensitivity (SE), precision (PR), Mathew's Correlation Coefficient (MCC) and Balanced Acurracy (BACC). The formulas for these metrics are expressed as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

$$SE = \frac{TP}{TP + FN} \quad (4)$$

$$PR = \frac{TP}{TP + FP} \quad (5)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

$$BACC = \frac{1}{2}[SE + SP] \quad (7)$$

where TP, FP, TN and FN are true positive, false positive, true negative and false negative, respectively.

Results and Discussion

A BACE-1 inhibitory activity classifier was built using a dataset of 1522 compounds (50% original samples, 50% GAN generated) and the Support Vector Machine (SVM) as the model building technique. This model comprised of 16 variables selected employing the GA-based feature selection procedure for the training set and was rigorously validated following a 10-fold

external validation procedure yielding good classification parameters, *i.e.* ACC = 0.91, SE = 0.90, SP = 0.91, PR = 0.89, MCC = 0.81 (Table 1, test set parameters). Subsequently, the built classifier was evaluated over the other 50% of the original samples and satisfactory predictivity was obtained, i.e. ACC = 0.77, SE = 0.83, SP = 0.73, PR = 0.72, MCC = 0.56 (Table 1, evaluation set).

In order to assess the effect of using GAN generated samples on the quality of the obtained classifier, the GAN samples were removed from the dataset, and the SVM classifier fit using the same 16 variables and configuration (Table 1). Likewise, this classifier's performance was assessed over the evaluation set. As may be observed in Table 1, the 10-fold test set performance of the classifier built using the BACE samples exclusively was lower (ACC = 0.78, SE = 0.78, SP = 0.79, Pr. = 0.77, MCC = 0.57) when compared to that of the BACE+GAN classifier, but this improvement could be attributed to the contribution of the GAN samples randomly included in the test samples. On the other hand, the comparable performance is observed when the BACE+GAN and BACE classifier are assessed over the evaluation set (*i.e.* BACE+GAN/BACE, ACC = 0.77/0.76, SE = 0.83/0.78, SP = 0.73/0.74, PR = 0.72/0.72, MCC = 0.56/0.52).

Table 1. Test and evaluation sets performance parameters for built SVM classifiers for BACE + GAN and the BACE datasets, respectively.†

Dataset	Dataset Splitting	ACC	SE	SP	PR	MCC
BACE+GAN/SVM	Test (25%)$^{\delta}$	0.91	0.90	0.91	0.89	0.81
	Evaluation (50%)	0.77	0.83	0.73	0.72	0.56
BACE/SVM	Test (25%)$^{\gamma}$	0.78	0.78	0.79	0.77	0.57
	Evaluation (50%)	0.76	0.78	0.74	0.72	0.52
BACE+GAN/RF	Test (25%)$^{\delta}$	0.89	0.90	0.87	0.86	0.78
	Evaluation (20%)	0.80	0.81	0.80	0.77	0.59
ECFP6/Bayes[8]	Evaluation (20%)	0.76	0.87	0.67	-	0.54
Canvas/RF[8]	Evaluation (20%)	0.81	0.77	0.84	-	0.61
Canvas/DNN[8]	Evaluation (20%)	0.82	0.75	0.88	-	0.64

†SVM: optimum C and gamma values determined by GridSearchCV algorithm; ‡average values from 10-fold external validation; δpercentages are with respect to the 50% of original data+GAN samples; γ percentages with respect to 50% of original dataset, exclusively

We also sought to assess the explore the chemical space mapped by the GAN relative to the BACE dataset. To this end, principal component analysis (PCA) was performed on the BACE+GAN data matrix and the components 1 and 2 (*i.e.* PC1 and PC2), explaining 54% and 18% of the dataset variance, respectively, were plotted using pyplot framework available in the matplotlib module. The PCA technique allows to represent multivariate data with a reduced number of orthogonal variables (principal components) such that PC1 explains the most variance of the original dataset, PC2 attempts to explain the remaining variance not explained by PC1, and so on. Figure 1 is a score plot of the principal components 1 and 2 for the BACE+GAN dataset. As anticipated, both the GAN and BACE samples are strongly loaded in PC1 and PC2 However, it is evident that they possess dissimilar weights which suggests that the former are not simply bootstrapped samples (duplications) of the latter but rather a possess their own identity in the BACE-1 inhibitory activity space.

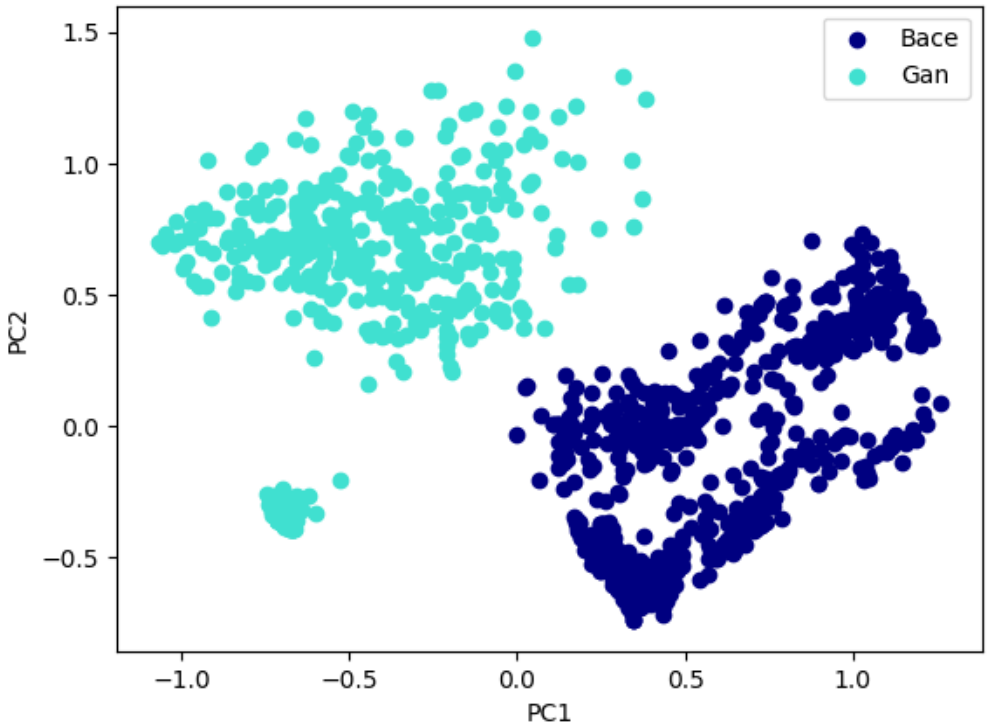

Figure 1. Score plot of principal components 1 and 2 for the BACE+GAN dataset.

We also compared the performance of the BACE+GAN classifier with other models reported in the literature using the same BACE dataset.[8] Considering that for these models the training and evaluation sets comprised of 80% and 20% of the dataset, respectively, the same splitting percentages were employed. Next, a GAN was employed to populate the training set yielding a total of 1988 compounds. A Random Forest (RF) classifier was built using the BACE+GAN training set and posteriorly employed to predict the activity profile of the evaluation set. It is interesting to note that BACE+GAN classifier provides better evaluation set parameters relative to the Bayesian model and similar performance to the RF and Deep Neural Networks (DNN) models (Table 1). Note that contrary to the BACE+GAN classifier built using 28 features, the Bayesian, RF and DNN models employed entire sets of variables, *i.e.* extended connectivity fingerprint (ECPF6) or Canvas descriptors since no feature selection procedures were applied. The obtained results suggest that GAN yields samples that adequately map the modeled chemical structure space and possibly allowing for improved prediction of molecular property profiles.

One of the possible utilities of GAN synthetic data is in balancing chemical compound datasets containing more samples in one class relative to the other(s). In this experiment, a GAN was employed to generate 901 synthetic active samples for the DENV inhibitory activity and consequently yielding a balanced dataset of 2119 compounds. An SVM classifier was built for the training set and validated over the test sets following a 10-fold external validation procedure. Table 2 shows the average test set performance of the built classifier (GAN samples were not included in the computation of these parameters), and a comparison with a DENV inhibitory activity model obtained using the SMOTE oversampling algorithm, as well as other classifiers reported in the literature, based on the random, repeated edited nearest neighbors, all kNN and

one sided selection undersampling algorithms.[9] The repeated edited nearest neighbors, all kNN and one sided selection are supervised undersampling strategies, used to progressively remove wrongly classified examples and thus yielding non-overlapping distributions.

Table 2. Comparison of test set performance of the DENV+GAN SVM-based classifier and with other models reported in the literature based on random, repeated edited nearest neighbors, all kNN, one sided selection undersampling algorithms, respectively.‡

Classification Models†	Balanced Acc.	MCC	Precision
DENV+GAN	0.81	0.70	0.83
SMOTE	0.74	0.61	0.86
Random[9]	0.85	0.57	0.49
Repeated Edited Nearest Neighbors[9]	0.84	0.68	0.73
All KNN[9]	0.84	0.68	0.72
One Sided Selection[9]	0.84	0.72	0.83

†Optimum SVM model C and gamma values were 12.0 and 0.28 for DENV+GAN, and 2.0 and 0.97 for the SMOTE-based classifier, respectively; ‡all models built using the same chemical compound dataset.

As may be observed the DENV+GAN model yields satisfactory 10-fold test set parameters with BACC = 0.81, MCC = 0.70 and PR = 0.83. In the comparison with the SMOTE-based classifier, it is observed that the DENV+GAN model yields better BACC (0.81 *vs* 0.74) and MCC (0.70 *vs* 0.61), while the former yields marginally better PR (0.83 *vs* 0.86). Moreover, the DENV+GAN model globally provides improved predictivity when compared to the classifiers obtained using the random (BACC = 0.85, MCC = 0.57, PR = 0.49), repeated edited nearest neighbors (BACC = 0.84, MCC = 0.63, PR = 0.73) and all KNN (BACC = 0.84, MCC = 0.68, PR = 0.72) undersampling algorithms, respectively, while comparable statistics are obtained relative to the one sided selection undersampling algorithm (BACC = 0.84, MCC = 0.72, PR = 0.83). However, it should be noted that the undersampling-based models are in fact majority vote-based ensembles built using 14 SVM base classifiers, and therefore are inherently more complex when compared to the single classifier BACE+GAN model.

In light of the satisfactory performance obtained with the BACE+GAN classifier, we believe that GANs may be useful for balancing chemical compound datasets containing less examples in one class relative to the other(s). Nonetheless it is important to note that GANs may not be used for balancing datasets with very few samples since just like any neural network model, GANs require to be trained with sufficient examples before they are employed to generate new examples for model building.

Conclusions

The mapping of the chemical structural space corresponding to a particular molecular property profile provides an opportunity to generate new data representative of this space without knowledge of the molecular structures. Such synthetic data may be useful for modeling tasks where additional data samples are needed to deal with data scarcity or class imbalance. In the present study, we evaluate the possible utility of GAN generated examples in model building envisioning 2 typical scenarios a) data scarcity: we populate the BACE-1 inhibitory activity data matrix using GAN synthetic examples and employ the expanded set for model building b) class imbalance: we generate synthetic examples to balance the DENV inhibitory activity dataset and subsequently build a classification model for this bioactivity endpoint. Interestingly, in both cases the obtained classifiers demonstrate satisfactory predictivity and compare favorably to other models reported in the literature for these endpoints. However, bearing in mind the computational complexity of the GAN-based approach and that no significant improvements are observed when GAN samples are used to populate balanced datasets, it is probably not beneficial for such cases. On the other hand, dataset balancing with GAN synthetic samples seem to outperform the popular oversampling approach SMOTE, although they do not improve the

results previously obtained with one sided selection undersampling. Even then, the authors are mindful that the performance of oversampling and undersampling techniques varies depending on the datasets and modeling techniques employed, as was previously demonstrated in the literature.[11] This is evocative of the "No Free Lunch Theorem" which reminds us that no single algorithm is a panacea to modeling problems.[12] We hope that GANs will be embraced by the chemoinformatics community as a promising tool to solve data imbalance and/or to enrich chemical data matrices for robust model building.

Associated Content

Supporting Information. Data matrices employed in the BACE-1 and DENV inhibitory activity modeling, as well as the python scripts employed to create the GAN synthetic samples and to build the reported classification models. This material is available free of charge via the Internet at

References

[1] A. Lavecchia, *Drug Discov. Today* **2015**, *20*, 318-331.
[2] aG. B. Goh, N. O. Hodas, A. Vishnu, *J. Comp. Chem.* **2017**, *38*, 1291-1307; bI. I. Baskin, V. A. Palyulin, N. S. Zefirov, in *Artificial Neural Networks*, Springer, **2006**, pp. 133-154; cI. I. Baskin, D. Winkler, I. V. Tetko, *Expert Opin. Drug Discov* **2016**, *11*, 785-795; dN. M. Halberstam, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Russ. Chem. Rev* **2003**, *72*, 629-649.
[3] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436-444.
[4] aH. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241-1250; bX.-W. Chen, X. Lin, *IEEE access* **2014**, *2*, 514-525; cE. Gawehn, J. A. Hiss, G. Schneider, *Mol. Inform.* **2016**, *35*, 3-14.
[5] aE. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, *J. Chem. Info. Model.* **2018**, *58*, 1194-1204; bM. Skalic, D. Sabbadin, B. Sattarov, S. Sciabola, G. De Fabritiis, *Mol. Pharmaceutics* **2019**, *16*, 4282-4291; cN. Anand, P. Huang, in *Adv. Neural Inf. Process Syst.*, **2018**, pp. 7494-7505.

[6] aX. Yi, E. Walia, P. Babyn, *Med. Image Anal.* **2019**, 101552; bL. Zhu, Y. Chen, P. Ghamisi, J. A. Benediktsson, *IEEE Trans. Geosci. Remote. Sens.* **2018**, *56*, 5046-5063.
[7] aS. Barua, M. M. Islam, X. Yao, K. Murase, *IEEE Trans. Know. Data Eng.* **2012**, *26*, 405-425; bA. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, *Learning from imbalanced data sets*, Springer, **2018**.
[8] G. Subramanian, B. Ramsundar, V. Pande, R. A. Denny, *J. Chem. Info. Model.* **2016**, *56*, 1936-1949.
[9] S. J. Barigye, J. M. G. de la Vega, J. A. Castillo-Garit, *J. Comput. Aided Mol. Des.* **2019**, *33*, 997-1008.
[10] aS. J. Barigye, Y. Marrero-Ponce, Y. M. López, O. M. Santiago, F. Torrens, R. G. Domenech, J. Galvez, *SAR & QSAR Environ. Res.* **2012**, 3-34; bS. J. Barigye, Y. Marrero-Ponce, V. Alfonso-Reguera, F. Pérez-Giménez, *Chem. Phys. Lett.* **2013**, *570*, 147-152; cS. J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, F. Torrens, L. M. Artiles-Martínez, R. W. Pino-Urias, O. Martínez-Santiago, *J. Comp. Chem.* **2013** *34*, 259-274; dS. J. Barigye, Y. Marrero-Ponce, O. M. Santiago, Y. M. López, F. Perez-Gimenez, F. Torrens, *Curr. Comput Aided Drug Des* **2013**, *9*, 164-183; eS. J. Barigye, Y. Marrero-Ponce, F. Pérez-Giménez, D. Bonchev, *Mol. Divers.* **2014**, *18*, 673-686.
[11] aA. Estabrooks, T. Jo, N. Japkowicz, *Comput. Intell.* **2004**, *20*, 18-36; bA. Gosain, S. Sardana, in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, **2017**, pp. 79-85.
[12] D. H. Wolpert, W. G. Macready, *IEEE T Evolut. Comput.* **1997**, *1*, 67-82.