Journal of Cheminformatics

COMMENT

Open Access

One size does not fit all: revising traditional paradigms for assessing accuracy of QSAR models used for virtual screening

James Wellnitz[1], Sankalp Jain[2], Joshua E. Hochuli[1], Travis Maxfield[1], Eugene N. Muratov[1*], Alexander Tropsha[1*] and Alexey V. Zakharov[2*]

Abstract

Traditional best practices for quantitative structure activity relationship (QSAR) modeling recommend dataset balancing and balanced accuracy (BA) as the key desired objective of model development. This study explores the value of the conventional norms in the context of using QSAR models for virtual screening of modern large and ultra-large chemical libraries. For this increasingly common task, we now recommend the use of models with the highest positive predictive value (PPV) built on imbalanced training sets as preferred virtual screening tools. This recommendation stems from practical considerations of how the results of virtual screening are used in experimental laboratories where only a small fraction of virtually screened molecules can be tested using standard well plates. As a proof of concept, we have developed QSAR models for five expansive datasets with different ratios of active and inactive molecules and compared model performance in virtual screening using BA, PPV, and other metrics. We show that training on imbalanced datasets achieves a hit rate at least 30% higher than using balanced datasets, and that the PPV metric captured this difference of performance with no parameter tuning. Importantly, hit rates were estimated for top scoring compounds organized in batches of the size of plates (for instance, 128 molecules) used in the experimental high throughput screening. Based on the results of our studies, we posit that QSAR models trained on imbalanced datasets with the highest PPV should be relied upon to identify and test hit compounds in early drug discovery studies.

Keywords Computer-assisted drug discovery, QSAR modeling, Imbalanced datasets, Virtual screening, Positive predictive value, Hit rate

Introduction

Quantitative structure–activity relationship (QSAR) modeling has been an integral part of computer-assisted drug discovery for over six decades [1]. This approach is used to rationalize the experimental data on chemical bioactivity measurements and develop models that can assess the expected bioactivity of new chemicals in advance of experiments. In addition to their broad use in drug discovery and chemical toxicity assessment [2, 3], methods and approaches used for QSAR modeling have proliferated into many areas of research [1]. These models are broadly categorized into two groups based on the type/format biological activity data modeled and overall

*Correspondence:
Eugene N. Muratov
murik@email.unc.edu
Alexander Tropsha
tropsha@email.unc.edu
Alexey V. Zakharov
alexey.zakharov@nih.gov
[1] Division of Chemical Biology and Medicinal Chemistry, Laboratory for Molecular Modeling,, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, USA
[2] National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, 9800 Medical Center Drive, Rockville, MD 20850, USA

goal of the model. Continuous models use raw quantitative data, like IC_{50} values, with the goal of also predicting the true value for a given compound. Classification models still utilize quantitative data, but in a discrete form where the goal is to predict which discrete 'class' a compound will be in. While it is possible to have many unique classes, it is common for QSAR classification models to be binary: categorizing compounds as either active or inactive. Each type of model employs specific metrics used for assessing its overall accuracy. Continuous models typically rely on cross-validation R^2, for accuracy assessment, while the accuracy of binary classification models is assessed using metrics derived from the confusion matrix, such as selectivity, specificity, and balanced accuracy (also known as Correct Classification Rate [4]). Achieving the highest balanced accuracy, i.e., models that can equally well predict both the positive *and* negative classes for the entire external set of molecules, is the most common assessment of binary classification model performance [5]. Further, when dealing with imbalanced classification datasets, down-sampling the predominant class is a common practice to enhance the balanced accuracy of the model [6].

The reliance on balanced accuracy as the key desired metric to characterize the accuracy of the binary classification QSAR models can be justified by the historical use of these models for lead optimization, where the aim has been to refine or design small sets of compounds to enhance the activity of a parent molecule [4]. Model section driven by balanced accuracy made sense when small training datasets of highly similar compounds and the recommended use of conservative applicability domains of QSAR models [7] resulted in a selection of the limited number of compounds from external libraries that was expected to include roughly the same ratio of active and inactive molecules as in the training sets. Furthermore, given the common imbalance in public datasets, which are skewed towards active molecules, and high-throughput screening (HTS) datasets, which are highly skewed towards inactive molecules, balancing these datasets through under-sampling the majority class (see, for instance, a recent rigorous study [8]) has been a conventional method prior to building models. These practices were also appropriate in the past when training sets were of limited size, virtual screening libraries have been relatively small, and the key context of use for QSAR models related to the task of hit or lead optimization. Thus, it is not surprising that popular best practices surrounding QSAR model development and validation have traditionally emphasized both the challenges associated with model development for imbalanced datasets [9] and the use of balanced accuracy as a key desired metric to characterize the performance of models [4].

The use of QSAR models is not just limited to lead/hit optimization, however. The possibility of using QSAR models for virtual screening and hit identification has been discussed in the literature as well [10] but practical utility of such applications has been limited by relatively small size of both training set and virtual screening libraries. Continuing rapid growth of expansive chemical bioactivity databases like ChEMBL [11] and PubChem [12], along with the exponential growth of make-on-demand chemical libraries [13] such as eMolecules Explore [14] and Enamines REAL Space [15], has significantly increased the appeal of using QSAR models as an alternative to structure based methods [16] in high throughput virtual screening (HTVS). In an HTVS campaign, these models can be used to screen ultra-large, multi-billion compound libraries; however, the ultimate practical objective is to nominate a small number of hit compounds for experimental validation [17]. Notably, false positive experimental nominations are expensive, both in terms of compound acquisition (synthesis or purchase) and the time and effort required to conduct the *in-vitro* and/or *in-vivo* experiments. The cost of experimental follow up also places a restriction on the number of compounds that can be selected for experimental validation, regardless of the size of the virtual screening libraries used. The result is only a small fraction of compounds that were screen (and potentially predicted active) being selected for experimental validation. These considerations underscore the critical importance of employing QSAR models that have high *positive predictive value* (PPV, often called precision) to nominate hit compounds for experimental testing [18–20]. Indeed, the high value of this metric calculated for the small selection of computational hit compounds implies high enrichment for active compounds, or, conversely, low rate of false positives, among the relatively small sets of nominated molecules. We demonstrated the success of the PPV-driven strategy for model building and virtual screening in a recent study that resulted in the discovery of novel binder of human angiotensin-converting enzyme 2 (ACE2) protein [18].

Here, we reconsider and revise previous best practices and recommendations driven by the considerations outlined above that in the modern age of medicinal chemistry and cheminformatics we face large biological screening datasets that are typically highly imbalanced in favor of inactive compounds for model training, and huge compound libraries employed for virtual screening, that are also expected to be even more imbalanced the same way. It is thus reasonable to acknowledge that both training and virtual screening sets are highly imbalanced, and different principles of building and assessing accuracy of QSAR models need to be considered when the goal is to

discover hits rather than optimize hits to leads. We posit that the PPV of a model is a better metric to assess performance at the virtual screening task: to enhance the proportion of active compounds identified in, by necessity, small selections of virtual screening hits as was also highlighted by Spiegel and Senderowitz [21]. We also emphasize the PPV of the highest ranked predictions as the best and most easily interpretable way to assess the expected performance of the model. We demonstrate on a case study of five HTS datasets that the common practice of balancing training sets to achieve models with high balanced accuracy is not optimal to address the primary goal of HTVS and results in models with a lower PPV, and worse HTVS performance, when utilized on external datasets. Our demonstration shows that balancing of the training sets, as expected, increases balanced accuracy while lowering the PPV, with imbalanced models having roughly 30% more true positives present in the top 128 predictions. We also compare to other metrics proposed for validation, including area under the receiver operating characteristic curve (AUROC) [22] and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) [23], and show that these metrics, while better than balanced accuracy, are not direct measures of virtual screening performance and can be difficult to interpret or parameterize. These findings strongly advocate for a paradigm shift in recommended best practices for constructing QSAR models to be used in HTVS of ultra-large libraries.

Results and discussion

Most HTVS campaigns often face constraints on the number of compounds that can be practically nominated for experimental testing. For instance, a model may predict 5000 compounds as putative actives, yet operational constraints often cap experimental testing. A quantitative HTS (qHTS) is limited to 128 compounds, corresponding to the throughput of a single plate in 1536 well format with 11 concentration points per compound. This necessitates a caveat in the model's objective to *nominate a top set of N compounds* with a minimal false positive rate [18]. In such scenarios, a model that identifies a smaller number of actives but ensures that such actives are included in the top, for example, 128 compounds are as, if not more, valuable as the one that can perfectly discriminate between active and in-active among all compounds screened. Conversely, a model that identifies 99% of all known actives in a large external set but includes none in the top 128 is ineffective for HTVS restricted by the size and number of plates used in the experimental screening. These considerations call to redefine the performance metrics, such that the number of actives within the top nominations of a fixed size emerges as a key indicator of a model's utility in HTVS.

The idea of virtual screening performance emphasizing the "high early enrichment" of actives among model predictions is not new and has been discussed nearly two decades ago by Truchon and Bayly [23] and more recently, by Speigel and Senderowitz [21]. Those works correctly identifies that other metrics, like area under the receiver operator curve (AUROC) and the enrichment factor (EF), are focused on assessing the ability to correctly classify active compounds *globally* across all predictions, rather than assessing just the performance *locally* on the top predictions which, as previously stated, is the true task of virtual screening. To address the lack of good metrics, Truchon and Bayly proposed a new one, the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC), which is an adjustment of AUROC meant to place additional emphasis on the performance of the top ranked predictions. However, BEDROC is characterized by an α parameter that can have a dramatic impact on the reported metric. Further, how to tune or select this parameter is not straight forward, as its impact on the resulting value is not linear nor easily interpretable. Overall, this can make it difficult to understand what the BEDROC value represents and how to interpret it in the context of model performance beyond "bigger is better". We suggest there is no need to use such a complex metric, as the commonly utilized PPV metric directly measures the model's ability to correctly identify actives. A simple adjustment to calculate the PPV on only the top $\boldsymbol{N}$ predictions is a direct measurement of how we would expect the model to perform when used for a virtual screening task that only allowed for $\boldsymbol{N}$ predictions.

To demonstrate this, we employed QSAR models in simulated virtual screening studies utilizing five distinct datasets (Table 1). In these datasets, generated from qHTS screenings, only a small fraction of molecules was active, which is a common outcome of large library screening campaigns. QSAR models were developed for training sets sampled from each set of assays in Table 1. Table 2 presents the different performance metric values for gradient boosting tree (GBT) models developed for five datasets, each analyzed under balanced and imbalanced training conditions. The datasets were evaluated for the number of active compounds correctly identified within the top 128 selections (referred to as PPV-128), alongside other performance metrics described above, including balanced accuracy (BA), AUROC, BEDROC [23] with varying alpha levels, and PPV (as a percentage rather than raw number of actives). Our findings indicate a significant increase (approx. twofold) in PPV-128 for models built with the original, imbalanced datasets,

Table 1 Size and prevalence of actives and in-actives of datasets used

Dataset names	AID: 504466	AID: 485314	AID: 485341	AID: 624202	AID: 651820
Total compounds	310,403	306,830	285,970	351,201	268,119
Num actives	4108	4348	1694	3902	10,727
Num in-actives	306,295	302,482	284,276	347,299	257,392
Ratio	1:75	1:70	1:168	1:89	1:24

AID refers to the PubChem Assay ID for this dataset. Datasets and table are from Zakharov et al. [24]

Table 2 Comparative validation/test set performance metrics of QSAR models on five PubChem datasets

ASSAY	Model type	PPV-128	BA	AUROC	BEDROC (α = 20)	BEDROC (α = 100)	PPV
aid_485314	Balanced	57.9 ± 5.7	0.80 ± 0.01	0.87 ± 0.01	0.22 ± 0.01	0.08 ± 0.01	0.06 ± 0.01
	Imbalanced	90.2 ± 4.6	0.60 ± 0.01	0.88 ± 0.01	0.27 ± 0.02	0.33 ± 0.02	0.70 ± 0.04
aid_485341	Balanced	5.9 ± 1.8	0.69 ± 0.01	0.75 ± 0.02	0.12 ± 0.01	0.03 ± 0.01	0.01 ± 0.01
	Imbalanced	12.1 ± 3.8	0.50 ± 0.01	0.75 ± 0.02	0.07 ± 0.01	0.03 ± 0.01	0.20 ± 0.11
aid_504466	Balanced	54.7 ± 4.1	0.83 ± 0.01	0.90 ± 0.01	0.25 ± 0.01	0.08 ± 0.01	0.06 ± 0.01
	Imbalanced	91.1 ± 4.0	0.60 ± 0.01	0.92 ± 0.01	0.26 ± 0.02	0.32 ± 0.02	0.75 ± 0.03
aid_624202	Balanced	19.6 ± 3.3	0.80 ± 0.01	0.88 ± 0.01	0.20 ± 0.01	0.06 ± 0.01	0.04 ± 0.01
	Imbalanced	40.3 ± 5.7	0.51 ± 0.01	0.89 ± 0.01	0.08 ± 0.01	0.06 ± 0.01	0.41 ± 0.08
aid_651820	Balanced	63.4 ± 5.3	0.78 ± 0.01	0.86 ± 0.01	0.24 ± 0.01	0.13 ± 0.01	0.13 ± 0.01
	Imbalanced	99.1 ± 4.2	0.58 ± 0.01	0.88 ± 0.01	0.28 ± 0.01	0.45 ± 0.01	0.66 ± 0.02

The value of each metric is calculated as the average of from 10 models. PPV-128 refers to number of actives in the top 128 predictions

emphasizing the advantage of such models for the use in HTVS campaigns. The observation that balancing has such a negative effect on the virtual screening performance can be explained by the observation that models for balanced datasets are trained on the data with a label distribution that does not match the expected distribution during virtual screening (actives are far rarer than negatives). In this case, up-sampling techniques, like SMOTE [25] to balance datasets by adding more positive examples would have the same effect as down sampling the negatives. However, if the negative class is underrepresented in the training data (which was not the case in this study), using SMOTE or other methods [26, 27] to generate negative datapoints can help improve the model performance measured by PPV as observed in this work.

Using PPV-128 as a baseline for true model performance at the HTVS task, we evaluated the conventional QSAR model assessment metrics (Table 2). BA stood out as the worst, having an inverse relationship with the number of nominated actives. This outcome reinforces the assertion that, although BA is a widely recognized metric of QSAR model accuracy, it is not the most appropriate metric for evaluating the performance of the models as applied to nominating compounds for HTVS. AUROC demonstrated better performance, marginally favoring imbalanced models over balanced ones. However, it fails to truly highlight the scale at which one model outperforms the other at the HTVS nomination task, and its reliability diminished for two assays, aid_485341 and aid_624202. Notably, the BEDROC metric was originally proposed for assessing models at the virtual screening task. Yet, like AUROC, it failed to differentiate the more effective model for assays aid_485341 and aid_624202. Further, the effectiveness of BEDROC notably declined when the alpha was reduced to 20, incorrectly labeling the balanced models as better for 2 out of 5 datasets. This pattern suggests that BEDROC's performance heavily relies on carefully adjusting its alpha setting based on the dataset. The need for precise tuning makes BEDROC more complex and less user-friendly than other, more straightforward metrics. PPV, especially PPV-128 stands out as able to differentiate the better preforming imbalanced models on all datasets while requiring no tunable parameters and a simple calculation and interpretation. This effect was still observed even when conventional applicability domain filtering [28] was used on predictions, which appeared to have minimal effect on the overall performance (Table S1). Enrichment factor (EF) was also calculated for each model (Table S2) and was as capable as PPV in differentiating better models. This is unsurprising, as EF is simply the PPV normalized by the rate of positives in the datasets (in some definitions [23]). Thus, as this rate is constant for all models trained on the same datasets, it is simply a linear scaling of PPV.

Since a HTVS campaign is focused on a single target, and thus, dataset such normalization has no effect. Instead, it only makes it harder to directly interpret the meaning of the value in the context of the true positive rate; PPV represents the true positive rate directly, whereas EF represents how much better a model's true positive rate is over random. Further, after using the ML model to nominate compounds for experimental follow up, it would be impossible to measure the true EF of the model on this set, as it would require knowledge (or a good estimate) of the true rate of actives for *all* compounds in the chemical library. PPV can be calculated using only the compounds from the nominated set, making it far easier to compare *in-silico* prediction to *in-vitro/vivo* experimental data, a crucial assessment to make for any HTVS campaign.

Similar trends were observed when using two different model algorithms: Random Forest (Table S3) and Deep Learning Dense Network (DL) classifiers (Table S4). Both had relatively similar metrics overall to the GBT models and showed a dramatic improvement in PPV and PPV-128 when using the full imbalanced dataset when compared to the balanced version.

Generalizing PPV-128

To generalize PPV-128, we can instead think of it as PPV-of-the-top-N ***(PPV-N)***, where N is the intended number of compounds desired for the experimental follow-up. This metric represents the proportion of active compounds within the nominated set relative to the *total number of nominations* rather than to the size of the test set. This makes it distinct from other early enrichment metrics, like EF or Robust Initial Enhancement (RIE) [29], that are dependent on the proportion of actives compounds in the training set. PPV-N is invariant to this proportion, since it only considers the small set of top N compounds and nothing else. This makes PPV-N directly pertinent to nomination-constrained HTVS campaigns, as synthesis and experimental characterization will only be pursued for the top N nominated compounds with the other prediction being ignored. EF-N is also invariant to the total number of nominations, however as discussed is a constant scalar transformation of PPV-N that convolutes the easy and direct interpretation that PPV-N has, which we assert makes it less useful, especially when both are easy to calculate. To better show the generalizability of PPV-N, we evaluated all models using this metric across a spectrum of N values, ranging from 16 to 1000. Our findings revealed that, with the exception of aid_485341, imbalanced models consistently surpassed their balanced counterparts in performance, particularly at $N=128$ and $N=256$ which correspond to the capacities of one or two 1536-well qHTS plates (Fig. 1). This observation further challenges the traditional QSAR modeling practices that advocate for dataset balancing, suggesting that such an approach may not confer an advantage in HTVS contexts.

Conclusions

Broadly utilizing a single metric, like balanced accuracy, as the best metrics for all QSAR models can impact the downstream performance of a model put into actual use, as is the case when models are used for virtual screening of large external libraries. This study thus recommends a shift in QSAR modeling practices for assessing model performance for HTVS applications toward using PPV, or more specifically PPV-N nominated compounds, instead of conventional BA. The results from five large, imbalanced, and diverse chemical datasets demonstrate that imbalanced models, which forgo the balancing of training sets, not only increase PPV but also significantly enhance the number of true positives within the top selections of virtual screening hits. We posit that using this metric will reduce false discoveries and associated costs in compound nomination. Our findings also highlight the shortcomings of BA as a performance metric when nominating compounds for HTVS, as it does not correlate with the identification of true actives, contrary to the objectives of HTVS.

As cheminformatics continues to evolve, it will continue to embrace methodologies that directly address the needs of modern drug discovery for more efficient and cost-effective screening processes. We posit that approaches considered and advocated for in this study may also find broader use in the domain of information retrieval, where early enrichment is often a primary goal [30], including fields as diverse as chemistry [31], medicine [32], genomics [33], and even music [34]. Thus, more robust approaches for assessing and building models for early enrichment tasks can result in benefits to all these areas.

Methods

We utilized five publicly available chemical datasets sourced from PubChem, as detailed by Zakharov et al. [24]. These datasets were characterized by binary class annotations for each compound, encompassing a diverse range of active to inactive ratios ranging from 1:24 to 1:189 (Table 1). Each dataset was split into training and validation sets randomly in an 80/20 ratio, respectively, using a stratified splitting schema to maintain a similar distribution of actives between the train and test set.

To explore the effects of data balancing, we maintained each dataset in its original imbalanced state and created balanced versions by randomly under-sampling the majority class to match the size of the minority class. This balancing process was exclusively applied to the training

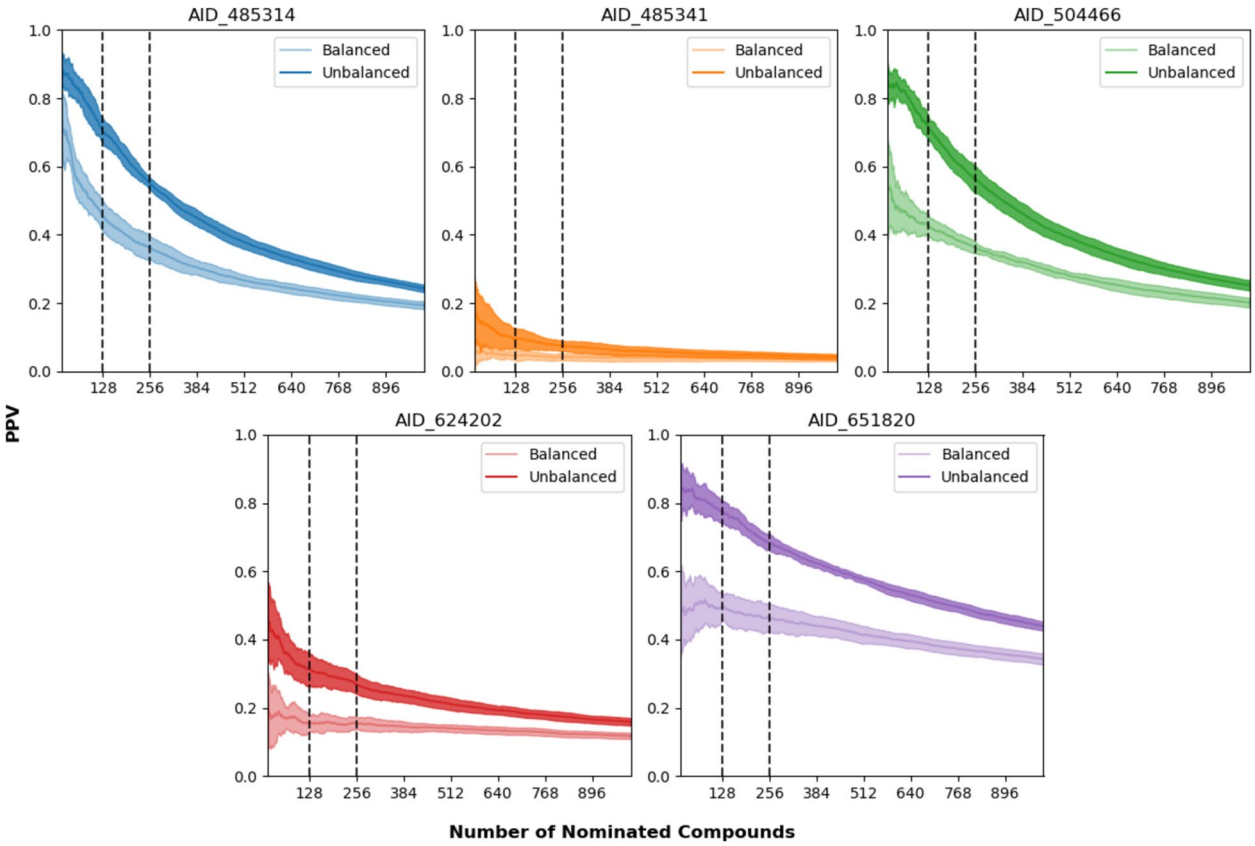

Fig. 1 Evaluation of PPV across varying nominations in HTVS campaigns

sets, while the validation sets preserved their original, imbalanced distributions, to mimic real world external libraries expected to include a small fraction of true actives.

For each dataset, we constructed a QSAR model using the gradient boosting method for both the balanced and imbalanced training sets. Default model settings were used and no hyperparameter, feature selection or model selection was used when training models; a single model was trained with all training data and then directly evaluated. Model performance was evaluated using several metrics including balanced accuracy (BA), positive predictive value (PPV, also known as precision), area under the receiver operator curve (AUROC), and the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC). For all discrete metrics (like BA), class membership was determined using a probability threshold of 0.5. These metrics were computed using the same external validation set for all models, ensuring no overlap between training and validation data points. To assess the robustness and variability of these metrics, we iteratively repeated this validation process 10 times per dataset, each time with a unique random split between training and test sets. This process was carried out using the StratifiedShuffleSplit (with n_fold set to 10) from SciKit-Learn [35]. All compounds were featurized using 2048-bit, radius 4 Extended Connectivity Fingerprints (ECPF) [36], computed via RDKit [37]. The gradient boosting models were developed with the XGBoost package [38] using default parameters. Random Forest models were built using default parameters from SciKitLearn. Neural networks models were built using 5 dense linear layers with hidden dimensions of 2000, 2000, 1000, 700, a rectified linear unit (ReLU) activation function, and the Adam optimizer with a learning rate of 0.0001, and trained with a binary cross entropy loss function until loss on a holdout validation set stopped decreasing. The validation set was a random 10% stratified split from the training dataset and was separate from the testing set used to evaluate the models. The network was implemented using Keras [39].

Applicability domain was calculated for test set compounds by finding the maximum Tanimoto similarity between the compound and the training set. If the Tanimoto similarity was above 0.35, the compound was considered in domain and its prediction kept.

For metrics that required a ranking of predictions (like AUROC or PPV-N) the classification model was asked to

provide a probability of a compound being in the "active" class. All models used are natively able to provide this probability value. When selecting only the top N compounds, the predictions with the highest probability were chosen. In the event of a tie, various medchem filters implemented in the STOPLIGHT program [40] for compounds in question would be used. In practice, none of our experiments resulted in the tie breaking, thus this step was unnecessary.

Abbreviations

QSAR	Quantitative structure activity relationship
BA	Balanced accuracy
PPV	Positive predictive value
HTS	High-throughput screening
HTVS	High-throughput virtual screening
AUROC	Area under the receiver operator curve
BEDROC	Boltzmann-enhanced discrimination of receiver operating characteristic
EF	Enrichment factor
RIE	Robust initial enhancement
ECFP	Extended connectivity fingerprints

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00948-y.

Supplementary Material 1.

Acknowledgements
We would like to thank the members of the Molecular Modeling Lab at the University of North Carolina Chapel Hill for their helpful discussion regarding this work.

Author contributions
Conceptualization and study design: AVZ, AT, ENM. Study implementation: JW, SJ, JEH, TM, AVZ. Manuscript writing: JW, JEH, SJ, TM, ENM, AVZ, AT.

Funding
This study was supported in part by NIH (grant R01GM140154), NSF (grant DMS2344256), and by the Intramural research program of the NCATS, NIH. JW and TM were supported by the National Institute of General Medical Sciences (Awards T32GM135122 and T32GM08633, respectively). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability
The code and datasets utilized in this study are made available at https://github.com/molecularmodelinglab/plate-ppv. We also make available a Sci-Kit Learn API compatible metric function for calculating the PPV-N.

Declarations

Competing interests
AT and ENM are co-founders of Predictive, LLC, which develops novel alternative methods and software for toxicity prediction. All other authors declare they have nothing to disclose.

Received: 15 November 2024 Accepted: 3 January 2025
Published online: 16 January 2025

References

1. Muratov EN, Bajorath J, Sheridan RP et al (2020) QSAR without borders. Chem Soc Rev 49:3525–3564. https://doi.org/10.1039/D0CS00098A
2. Siramshetty VB, Nguyen D-T, Martinez NJ et al (2020) Critical assessment of artificial intelligence methods for prediction of hERG channel inhibition in the "Big Data" Era. J Chem Inf Model 60:6007–6019. https://doi.org/10.1021/acs.jcim.0c00884
3. Jain S, Siramshetty VB, Alves VM et al (2021) Large-scale modeling of multispecies acute toxicity end points using consensus of multitask deep learning methods. J Chem Inf Model 61:653–663. https://doi.org/10.1021/acs.jcim.0c01164
4. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488. https://doi.org/10.1002/minf.201000061
5. Golbraikh A, Wang XS, Zhu H, Tropsha A (2016) Predictive QSAR modeling: methods and applications in drug discovery and chemical risk assessment. In: Leszczynski J (ed) Handbook of computational chemistry. Springer, Netherlands, pp 1–48
6. Golbraikh A, Shen M, Xiao Z et al (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253. https://doi.org/10.1023/A:1025386326946
7. Kar S, Roy K, Leszczynski J (2018) Applicability domain: a step toward confident predictions and decidability for QSAR modeling. In: Nicolotti O (ed) Computational toxicology: methods and protocols. Springer, New York, pp 141–169
8. Norinder U, Boyer S (2017) Binary classification of imbalanced datasets using conformal prediction. J Mol Graph Model 72:256–265. https://doi.org/10.1016/j.jmgm.2017.01.008
9. Casanova-Alvarez O, Morales-Helguera A, Cabrera-Pérez MÁ et al (2021) A novel automated framework for QSAR modeling of highly imbalanced leishmania high-throughput screening data. J Chem Inf Model 61:3213–3231. https://doi.org/10.1021/acs.jcim.0c01439
10. (2008) Chemoinformatics approaches to virtual screening. The Royal Society of Chemistry
11. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1093/nar/gkr777
12. Kim S, Chen J, Cheng T et al (2023) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380. https://doi.org/10.1093/nar/gkac956
13. Cherkasov A (2023) The 'Big Bang' of the chemical universe. Nat Chem Biol 19:667–668. https://doi.org/10.1038/s41589-022-01233-x
14. eMolecules Explore. http://www.emolecules.com/explore. Accessed 5 Mar 2023
15. REAL Space-Enamine. https://enamine.net/compound-collections/real-compounds/real-space-navigator. Accessed 5 Mar 2023
16. Bender BJ, Gahbauer S, Luttens A et al (2021) A practical guide to large-scale docking. Nat Protoc 16:4799–4832. https://doi.org/10.1038/s41596-021-00597-z
17. Neves BJ, Braga RC, Melo-Filho CC et al (2018) QSAR-based virtual screening: advances and applications in drug discovery. Front Pharmacol 9
18. Hochuli JE, Jain S, Melo-Filho C et al (2022) Allosteric binders of ACE2 are promising anti-SARS-CoV-2 agents. ACS Pharmacol Transl Sci 5:468–478. https://doi.org/10.1021/acsptsci.2c00049
19. Jain S, Talley DC, Baljinnyam B et al (2021) Hybrid in silico approach reveals novel inhibitors of multiple SARS-CoV-2 variants. ACS Pharmacol Transl Sci 4:1675–1688. https://doi.org/10.1021/acsptsci.1c00176
20. Abrams RPM, Yasgar A, Teramoto T et al (2020) Therapeutic candidates for the Zika virus identified by a high-throughput screen for Zika protease inhibitors. Proc Natl Acad Sci 117:31365–31375. https://doi.org/10.1073/pnas.2005463117
21. Spiegel J, Senderowitz H (2020) Evaluation of QSAR equations for virtual screening. Int J Mol Sci 21:7828. https://doi.org/10.3390/ijms21217828
22. Matveieva M, Polishchuk P (2021) Benchmarks for interpretation of QSAR models. J Cheminformatics 13:41. https://doi.org/10.1186/s13321-021-00519-x
23. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "Early Recognition" Problem. J Chem Inf Model 47:488–508. https://doi.org/10.1021/ci600426e
24. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC (2014) QSAR modeling of imbalanced high-throughput screening data in PubChem. J Chem Inf Model 54:705–712. https://doi.org/10.1021/ci400737s

25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
26. Kumari C, Abulaish M, Subbarao N (2020) Using SMOTE to deal with class-imbalance problem in bioactivity data to predict mTOR inhibitors. SN Comput Sci 1:150. https://doi.org/10.1007/s42979-020-00156-5
27. Cáceres EL, Mew NC, Keiser MJ (2020) Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. J Chem Inf Model 60:5957–5970. https://doi.org/10.1021/acs.jcim.0c00565
28. Sushko I, Novotarskyi S, Pandey A et al (2010) Applicability domain for classification problems. J Cheminformatics 2:P41. https://doi.org/10.1186/1758-2946-2-S1-P41
29. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J Chem Inf Comput Sci 41:1395–1406. https://doi.org/10.1021/ci0100144
30. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval, illustrated edn. Cambridge University Press, New York
31. Edgar SJ, Holliday JD, Willett P (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. J Mol Graph Model 18:343–357. https://doi.org/10.1016/S1093-3263(00)00061-9
32. Gupta D, Loane R, Gayen S, Demner-Fushman D (2022) Medical image retrieval via nearest neighbor search on pre-trained image features
33. Nadkarni PM (2002) An introduction to information retrieval: applications in genomics. Pharmacogenomics J 2:96–102. https://doi.org/10.1038/sj.tpj.6500084
34. Byrd D, Crawford T (2002) Problems of music information retrieval in the real world. Inf Process Manag 38:249–272. https://doi.org/10.1016/S0306-4573(01)00033-4
35. Kramer O (2016) Scikit-Learn. Machine learning for evolution strategies. Springer International Publishing, Cham, pp 45–53
36. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
37. Landrum G, Tosco P, Kelley B et al (2023) rdkit/rdkit: 2023_03_1 (Q1 2023) Release
38. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 785–794
39. Keras: Deep Learning for humans. https://keras.io/. Accessed 22 Dec 2024
40. Wellnitz J, Martin H-J, Anwar Hossain M et al (2024) STOPLIGHT: a hit scoring calculator. J Chem Inf Model 64:4387–4391. https://doi.org/10.1021/acs.jcim.4c00412

Publisher's Note