https://doi.org/10.48048/tis.2023.5388

Prediction of Cytotoxicity Against HepG2 by Quantitative Structure-Activity Relation (QSAR) Modelling

Nhung Phuong Nguyen[1,*], Oanh Kieu Thi Nguyen[2], Nghia Dinh Tran[1] and Hai The Pham[3]

[1]*Department of Basic Sciences, Hanoi University of Pharmacy, Hanoi 10000, Vietnam*
[2]*Department of Life Sciences, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, Hanoi 100000, Vietnam*
[3]*Department of Medicinal Chemistry, Hanoi University of Pharmacy, Hanoi 100000, Vietnam*

([*]Corresponding author's e-mail: nhungnguyen.hup.edu@gmail.com)

Received: 26 July 2022, Revised: 13 September 2022, Accepted: 20 September 2022, Published: 18 March 2023

Abstract

Hepatocellular carcinoma (HCC) is the dominant subtype of liver cancer with very low survival rate but the chemotherapy for HCC is still in grey zone due to the limited efficacy and high toxicity profile of approved drugs raising the heavy demand on drug development for HCC. The study aimed to establish a desirability based quantitative structure activity relation (QSAR) model to predict the activity of chemical compounds against one of HCC cell line (HepG2). Different support vector machine (SVM) models were constructed and ensembled to 10 virtual screening protocols. These protocols were validated by an external dataset in combination with decoys as interference. Results showed that ensemble models exhibited improved area under the Receiver Operating Characteristic Curve (ROC), sensitivity, and specificity compared to base models in training and test set. When being validated for virtual screening to recover known active molecules in the mixture with known inactive and decoy compounds, all virtual screening protocols have good performance with good Boltzmann-Enhanced Discrimination of ROC (BEDROC) and enrichment factor (EF). The best protocol with BEDROC of 0.63 and EF of 29.55 was suitable for further screening of active compounds against HepG2 cell line.

Keywords: QSAR, Cancer, Cytotoxicity, Desirability, Virtual screening, Machine learning, Model

Introduction

Hepatocellular carcinoma accounts for more than 80 % of liver cancer cases with very low 5 year survival rate [1]. The application of systemic therapy for treatment for HCC is extremely limited regarding to the effectiveness and toxicity of available approved drugs including kinase inhibitors and immune checkpoint inhibitors. Sorafenib, the initial 1st line approved drug for patient at advanced stages, is an expensive drug with high toxicity including diarrhea, palmar-plantar erythrodysesthesia and hemorrhage. Lenvatinib, another 1st line recommendation for non-excisable HCC, causes not only similar adverse events to sorafenib but also other serious side effects such on kidney, hematology disorder although this drug have been reported to increase more 4 months of survival rate than sorafenib. As the 2nd line treatment, antibodies such as nivolumab and pembrolizumab have been reported to have better safety profile but their efficacy has not been satisfactory in clinical trials. The current trend for systemic treatment of HCC is the combination of kinase inhibitors and immune check point inhibitors [1].

As traditional drug development approach with experimental testing has its own disadvantages of time and financial cost, computer aided study which serves as a preliminary tool or guidance for experimental testing, has been becoming more and more important in the whole process. For anticancer drug identification, numerous models have been constructed and successfully applied to understand the correlation between structure and activity as well as predict the desired properties by quantitative structure activity relation (QSAR). Apart from activity prediction, QSAR could be also able to generalize safety profile of a drug candidate and therefore could forecast the success of the candidate through later preclinical and clinical phases [2].

One of the reasons make HCC "fade" under the action of current chemotherapy is the unclear pathological pathways as well as unknown markers and essential proteins in the cancer network. Besides,

HCC has multiple phenotypes which limit the application scope of drugs. Therefore, the more priority of HCC drug development is seeking for a broad-spectrum compound against multiple cell lines and if possible, multiple targets. Multi-objective optimization has been proposed for this purpose and applied to several cancers as well as different objectives [3,4]. However, there is little multiple objective prediction research for liver cancer in general and HCC specifically [5].

In this study, as initial step of larger multi-objective project, our objective is to model the quantitative structure: Activity relation as a protocol for virtual screening of drug cytotoxicity against one of liver cancer cell line, HepG2 based on its availability for further *in vitro* testing when we apply the developed protocols to screen out several commercial molecule databases in future studies.

Materials and methods

Data curation

Data for cytotoxicity modelling were obtained from Chembl database (https://www.ebi.ac.uk/chembl/g) with keyword of HepG2 (CHEMBL395 cell line identification code). Then we filtered by following options: IC50, cell-based format, and organism *Homo sapiens*. The raw data were appended in Supplementary Information SI.0. From downloaded data, we initiated data curation by eliminating duplicate compound id, SMILE code, and stereoisomer molecules and only retaining minimum activity ones. Compounds with meaningless or [2H] beginning SMILE code were also pulled out. To increase the harmonization of experimental procedure between published data, we only selected compounds which were evaluated with cytotoxicity by MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) method in 48 h.

From SMILE code, we converted the representation of molecules to sdf format by OpenBabel tool [6], and through this conversion, we also performed almost standardization with deletion of hydrogen to make implicit, transformation of dative bonds to make homogenized presentation of conjugated groups such as NO_2, and stripping salt. The last step of standardization, e.g. aromatization, was done by Standardizer software (version 22.6.0, ChemAxon (http://www.chemaxon.com)). As a molecule could be presented by multiples SMILE codes, afterwards, we used ISIDA/EdiSDF tool to detect duplicate structures.

Molecular descriptors were taken from avalDesc tool. There are 477 descriptors of 6 groups (constitutional indices, ring descriptors, burden eigenvalues, functional group counts, molecular properties, atom centered fragments). The name and description of 477 descriptors are shown in Supplementary Information SI.1. Features were scaled in range 0 and 1. Then all constant or quasi constant features were removed.

Each molecule is considered as a vector of standardized molecular descriptors. Then we detected activity cliffs which have similar structure but significant difference of activity by calculate 2 corresponding matrices: Similarity and potency difference matrix [7]. Similarity ($NormED_{ij}$) was calculated by the Euclidean distance between pair of molecules (ED_{ij}) and then also normalized in range 0 and 1 by dividing to the maximum distance ($maxED_{ij}$):

$$NormED_{ij} = \frac{ED_{ij}}{maxED_{ij}} \quad (1)$$

Potency difference (ΔP_{ij}) was the absolute difference between logarithm of IC_{50} value in μM unit of 2 molecules:

$$\Delta P_{ij} = \left|P_i - P_j\right| \quad (2)$$

An activity cliff is defined with molecular pair which has the potency difference of larger than 1 order of magnitude ($\Delta P_{ij} \geq 1.0$) and structure similarity of more than 90 % ($NormED_{ij} \leq 0.01$).

Subsequently, we defined IC50 threshold of 10 μM to categorize chemicals into 2 classes: Active and inactive. Dataset were divided into 3 sets: External, training, and test set. External set was generated by randomly sampling from initial curated dataset with 15 % of molecules. 80 % of the remaining molecules were categorized in training set and 20 % of remaining set were included in test set. The training and test set were sampled with the same ratio of 2 classes.

Base model construction

Initially, different base models are constructed to correlate the desired activity and structure parameters. In each model, we chose randomly a number of descriptors (5, 10 and 15), then constructed the model by support vector machine (SVM) radial algorithm with tuned C and sigma parameter (3 repeat and 10 fold cross validation) [8].

As applicability domain (AD) is crucial to validate a QSAR model as per OECD principles, the AD for our models is determined by distance based method [9]. Any molecule is considered as in AD if:

$$d \leq d_k + Z.s_k \tag{3}$$

in which d is the Euclidean distance of the molecule to the centroid of training dataset, d_k is the mean of the Euclidean distance of the molecule to the centroid of training dataset, s_k is the standard deviation of Euclidean distances of the molecule to the centroid of training dataset, and Z is the coefficient [9]. In our study, we assigned Z equal to 1.0 to detect the outliers.

Based on AD definition, outliers of 2 test sets were filtered. To evaluate the base models for ensemble in the next step, a model should have ROC value beyond a threshold (0.70) when predicting on test and training set [8].

Model ensemble

The risk of wrong decision to categorize of a molecule into one or another class could be neutralized by the average of the multiple model prediction [1,2]. Initially, for each cell line, we assessed whether a specific molecule was in AD of component models. When the AD criteria were satisfactory, the outputs of active class of the models were averaged.

The model selection was performed by genetic algorithm (GA) strategy with random forest as learning method, 50 search iterations and 2 evaluated subsets per iteration. The crossover and mutation probability were respectively 0.7 and 0.3. The performance of GA models was assessed by 10-fold cross validation.

Desirability score conversion

The aggregated score predicted by ensemble models are then converted to desirability score (D_i) by the Eq. (4) [2]:

$$D_i = \left[\frac{\widehat{S_i} - \min(S)}{\max(S) - \min(S)}\right]^s \tag{4}$$

which $\widehat{S_i}$ is the aggregated score of sample i, $\min(S)$ and $\max(S)$ are the minimum and maximum of aggregated scores across the training and 2 test sets, s is the exponential coefficient calculated by the formula:

$$0.5 = \left[\frac{|\min(S) - 0.5|}{\max(S) - \min(S)}\right]^s \tag{5}$$

For a new compound, the desirability number is:

$$D_i = \begin{cases} 0 \text{ if } \widehat{S_i} \leq \min(S) \\ \left[\frac{\widehat{S_i} - \min(S)}{\max(S) - \min(S)}\right]^s \text{ if } \min(S) < \widehat{S_i} < \max(S) \\ 1 \text{ if } \widehat{S_i} \geq \max(S) \end{cases} \tag{6}$$

which $\widehat{S_i}$ is the average score of the molecule.

Virtual screening validation

Molecules with known effect in external dataset were sorted by their desirability scores. We submitted the smile code of known active compound in the external dataset to DUD-E database (http://dude.docking.org/generate) to generate decoys for validation [3]. The performance of VS protocol was determined by following parameters: Area Under the Accumulation Curve (AUAC); ROC; Enrichment Factor (EF) and Boltzmann-Enhanced Discrimination of ROC (BEDROC) [8].

Suppose that we have N ranked compounds in which the number and ratio of active compounds are n and R_a respectively. x_i is the relative ranking of each compound.

The calculation of AUAC and ROC was based on following equations:

$$AUC = 1 - \frac{1}{n}\sum_{i=1}^{n} x_i \quad (7)$$

$$ROC = \frac{AUC}{R_i} - \frac{R_a}{2R_i} \quad (8)$$

$$R_i = 1 - R_a \quad (9)$$

ROC represents the ability of a model to recognize (or rank) an active before an inactive molecule. ROC has value in range of 0 and 1 and if the model ranking is not better than random ranking, ROC is equal to 0.5 [4]. AUAC is also an area under the curve value but it is calculated from an empirical curve which establishes the relationship between the counts of active molecules accumulated to a certain relative rank when molecules were ranked from worst to best position versus their relative ranks [12]. These 2 metrics are dependent on R_a and do not depict the early recognition effectiveness of the model [12].

Meanwhile, EF correlates to the ratio of active molecules identified in the fraction of $0 < \chi < 1$ of the ordered list relative to what is expected from a uniform ranking for that same fraction.

$$EF = \sum_{i=1}^{n} \frac{\delta_i}{\chi n} \quad (10)$$

where: $\delta_i = \begin{cases} 1 & x_i < \chi \\ 0 & x_i > \chi \end{cases}$

Although EF could well answer the question of how many active compounds are identified in the selected fraction compared to randomly sampling, the value could not discriminate various models with the same number of enriched molecules but different ranking position profiles of active compounds within the fraction. The discrimination of such models are implied by another metric, BEDROC. BEDROC is the probability that an active ranked by the evaluated method will be found before a compound that would come from a hypothetical exponential PDF with parameter R, or f(x) [12]. is calculated through Robust Initial Enhancement metric (RIE) as following equations [8]:

$$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{1}{N}\left(\frac{1-e^{-\alpha}}{e^{\frac{\alpha}{N}}-1}\right)} \quad (11)$$

$$RIE_{min} = \frac{1-e^{-\alpha R_a}}{R_a(1-e^{\alpha})} \quad (12)$$

$$RIE_{max} = \frac{1-e^{-\alpha R_a}}{R_a(1-e^{-\alpha})} \quad (13)$$

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \quad (14)$$

$$\theta(1 - e^{\alpha}) - 1 + e^{-\alpha z} = 0 \quad (15)$$

The significance of α is that across the active compounds, the ones at the top-ranking list have higher weights than those at the tail and contributes more to RIE. z is the percent of the ranked list at which enrichment is important and θ is the expected contribution of the enrichment at this z % fraction to the overall enrichment [2].

Results and discussion

Data curation

From Chembl database, we collected more than 21,000 compounds. Due to the inhomogeneity of the data source regarding to experimental procedure and result representative approach, data curation is extremely essential to avoid noise and interruption to QSAR modelling. We only retained 14,525 molecules with exact reported IC50 value and defined unit. Next, we selected the data of *Homo sapiens* with 14,491 compounds. Because 1 molecule could be studied in various research leading to different IC50 values, we firstly identified duplicate through Chembl identification code and smile code of substances and kept the minimum value of duplicates resulting the dataset of 12,374 molecules. Among these molecules, we continued to classify and choose 2,031 compounds which have evaluated with MTT cytotoxicity for 48 h. The structure of 2,031 molecules was transformed into sdf format and standardized before submitting to EditSDF tool to detect duplicate. There are 6 pairs of duplicates detected by the tool and we removed duplicates which have larger IC50 than the other. Finally, to increase the continuity of QSAR model, we used similarity matrix of 2,025 compounds based on their 466 descriptors and removed the cliffs to obtain the final input dataset of 838 molecules.

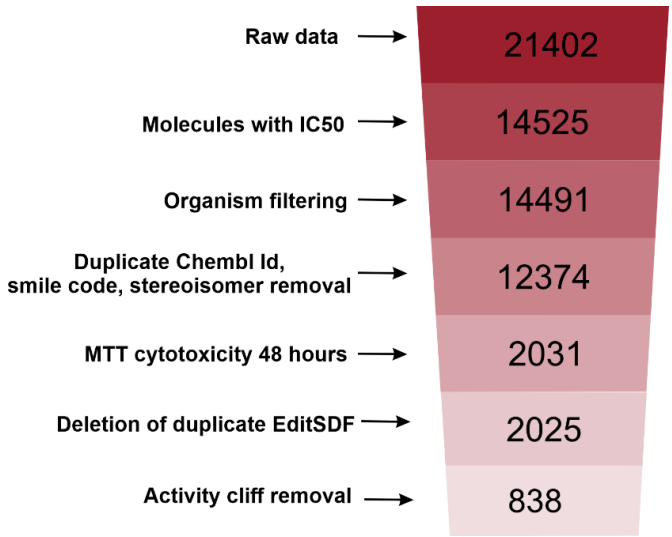

Figure 1 Number of molecules remaining after each step of data curation.

The standardized dataset contains 269 active (32.10 %) and 569 inactive (67.90 %) compounds. Afterwards, this dataset was divided into different dataset: external, training, and test set (**Table 1**). External dataset was used to validate entire modelling for virtual screening, so this was firstly sampled by random sampling with 126 molecules and 34.92 % of active molecules (44 molecules). The training set includes 570 molecules with 31.58 % of active compounds while test set has 142 molecules with 31.69 % of active compounds. The ratio of 2 classes in training and test set was not significantly imbalanced and expected to not have substantial effect on the performance of SVM model.

Table 1 Data splitting.

Dataset	Training	Test	External
Number of molecules	570	142	126
Number of active compounds	180	45	44
Ratio of active compounds (%)	31.58	31.69	34.92

Base models

Feature selection has strong impact on performance and predictivity of a QSAR model. From initial 477 molecular descriptors, 85 descriptors with missing and nearly unchanged values were moved out. Then, after data splitting, feature selection was proceeded with training dataset by deletion of highly correlated features with correlation threshold of 0.7. As our QSAR models were constructed by random sampling of

features, the obtained set of 154 molecular descriptors generated excessive numbers of possibility. Therefore, we did a preliminary SVM model with 154 descriptors and ranked the contribution of all descriptors to the performance of the big model. The name and description of 154 descriptors are highlighted (in red) in Supplementary Information SI.1. From the feature importance ranking, we selected 50 descriptors to build QSAR base models.

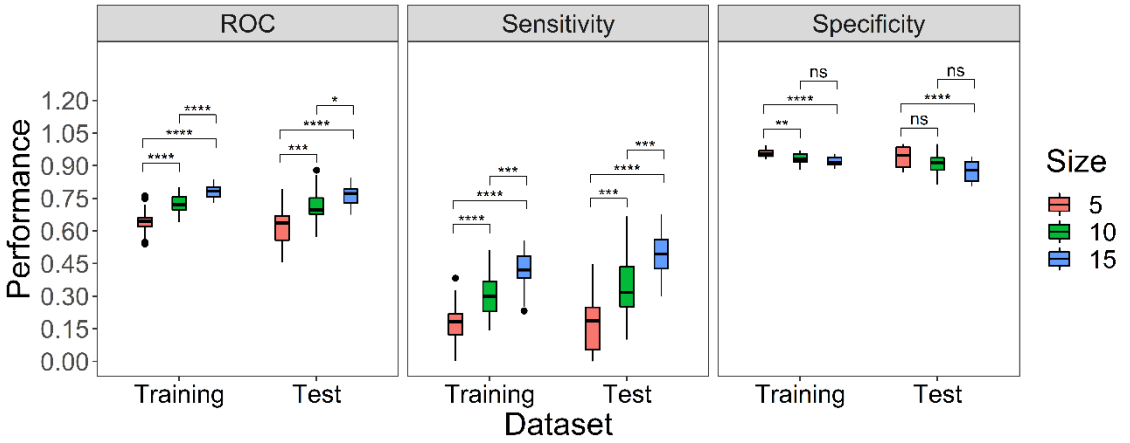

Figure 2 Performance of base models.

For base model, we selected randomly a number of descriptors (5, 10, or 15) and repeated 30 times for each size so that in total, there are 150 base models and their performances were represented in **Table 2** and **Figure 2**). Generally, base models have acceptable ROC and excellent specificity in both training and test set but the sensitivity values are low with maximum of less than 0.70 and some models have bad sensitivity of 0. There is no significant difference referring to all 3 criteria between training and test across 30 models. Referring to the impact of feature set size, ROC, and sensitivity of base models in both training and test set are improved but the sensitivity appears to be not significant difference and even decrease especially for test set when increasing the feature subset size. Despite of the good specificity, ROC of several base models is not satisfactory and the sensitivity extremely lower than expected.

Table 2 Summary of performance of base models.

Performance	Training set Mean ± standard deviation (min − max)	Test set Mean ± standard deviation (min − max)
ROC	0.72 ± 0.07 (0.54 – 0.84)	0.69 ± 0.10 (0.46 – 0.88)
Specificity	0.94 ± 0.03 (0.88 – 0.99)	0.91 ± 0.05 (0.80 – 1.00)
Sensitivity	0.30 ± 0.14 (0.00 – 0.55)	0.34 ± 0.18 (0.00 – 0.68)

Ensemble models

Choosing a single QSAR model as the best generalization of the function of activity and structure is often challenging because the outstanding of performance in training and test set does not guarantee the predictive power of the model especially when the training and test set is not the good representative of the predicted data. Besides, as the predictivity scope of each model is limited by its applicability domain, the decision making to assign a sample as one or another class is sometimes outside of the range or even the sample in applicability domain of the model, the boundary between 2 classes could be too complicated to clearly define the correct class. Therefore, integration of multiple models could not only neutralize the risk of choosing the "bad" single model but also expand the applicability domain and facilitate the definition of class boundary [11]. By averaging the probability of each molecule to be in active class, we ensembled base models which have ROC of above 0.70 and searched for model combination by genetic algorithm. The comparison of ensembles and their base models is illustrated in **Figures 3** and **4** and **Table 3**, respectively.

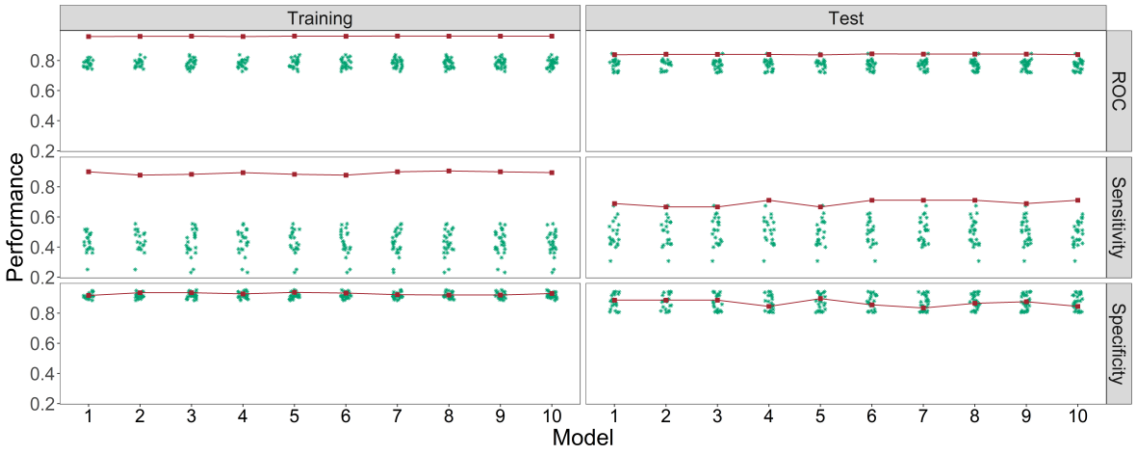

Figure 3 Comparison of performance of ensemble models to their component base models. ROC, sensitivity, and specificity of 10 ensemble and their corresponding component models are represented in red and green, respectively.

Table 3 Summary of performance of ensemble models.

Performance	Training set Mean ± standard deviation (min – max)	Test set Mean ± standard deviation (min – max)
ROC	0.96 ± 0.00 (0.96 – 0.96)	0.84 ± 0.00 (0.84 – 0.84)
Specificity	0.93 ± 0.01 (0.88 – 0.99)	0.87 ± 0.02 (0.84 – 0.90)
Sensitivity	0.89 ± 0.01 (0.88 – 0.91)	0.69 ± 0.02 (0.67 – 0.71)

It is clearly shown that ensemble models remarkedly enhance the performance of their component models. ROC of training set is improved better than that of test set. ROC of component models is in range of 0.75 and 0.85 for both test and training set. ROC of integrated models increases to 0.96 and 0.84 for training and test set respectively and this metric is higher than all component models. The sensitivity of base models is very low of under 0.65 but ensemble model improves significantly especially for training set of more than 60 % or even 100 % for some models. For test set, the sensitivity of ensemble models is higher or equal to the sensitivity of the best component models and in average, it is improved by around 30 %. The specificity of ensemble models is approximately 0.90 and higher than the majority of their components but still lower than several best models however with the comparison to the mean of base models, the specificity of ensemble is comparable to base models. The outperformance of ensemble models compared to the individual models is also evidenced in other studies [13,14].

As all ensemble models are equivalent to each other, we applied all 10 models for virtual screening to determine their ability to filter out the active molecules in a mixture of active, inactive and decoy.

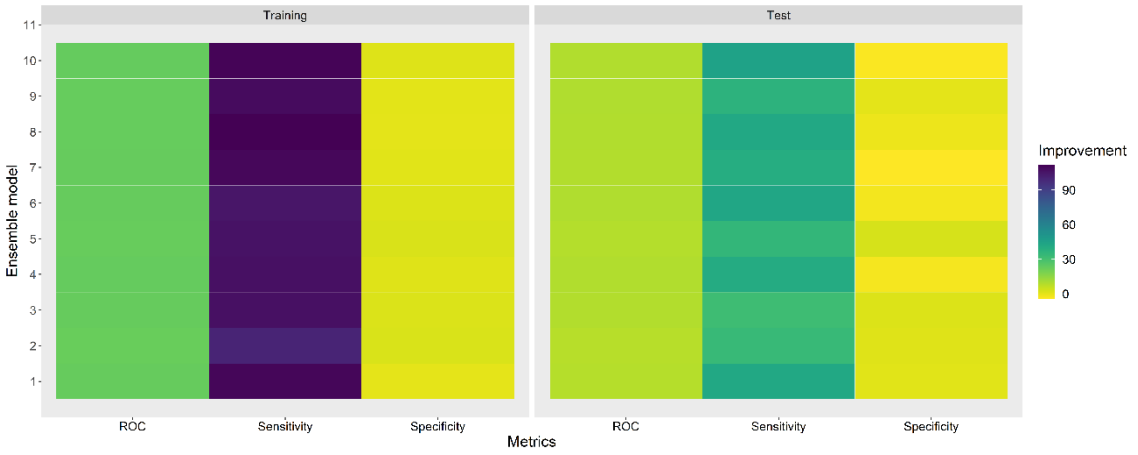

Figure 4 Relative improvement of ensemble models compared to the mean of component models. The relative improvement is the difference between the metric of each ensemble model the metric average of its component models divided by the metric average of its component models. The values are scaled from 0 (yellow) to 100 (purple).

Virtual screening

External dataset has 126 molecules with 44 active compounds. From these active molecules, we generated 2,107 decoys from DUDE database so that the ratio of active compounds (R_a) is 2.03 %. The protocol of base and ensemble models was applied to calculate the desirability score of decoys and molecules in external dataset and rank the molecules based on their scores. Each model was validated by the ability to early recognize meaning ranking the active molecules at the top of the list through BEDROC value. Besides, a perfect model would have to concentrate as many active compounds as possible in the selected fraction (e.g. top 1 %) by EF value [3]. Validation result of 10 models is shown in **Table 3**.

Table 4 Validation of VS screening protocol.

Model	AUAC	EF	ROC	BEDROC
1	0.72	29.55	0.72	0.63
3	0.71	29.55	0.72	0.62
6	0.71	29.55	0.71	0.62
2	0.70	29.55	0.71	0.61
10	0.70	29.55	0.70	0.60
8	0.70	29.55	0.71	0.60
9	0.69	27.27	0.69	0.60
5	0.70	29.55	0.70	0.60
4	0.70	27.27	0.71	0.59
7	0.68	25.00	0.68	0.54

All models have good BEDROC of above 0.5 and EF from 25 to 29.55. The performance of 10 models is comparable but model 1 exhibit the best score regarding to all metrics (AUAC, EF, ROC and BEDROC). **Table 5** shows the specific ranking of molecules from external dataset and DUDE generated decoys. Molecules with ChEMBL ID are from external dataset and their classes are known from their IC50 values from ChEMBL while decoys are considered as inactive class. Top 1 % fraction has 22 molecules in total, first 2 positions are of known active compounds. There is only one known inactive compound at 3rd ranking and the model recovers 14 of 44 known active compounds in the fraction so the real active compounds account for 63.64 % of the list.

Table 5 Detailed ranking of top molecules by model 1.

Name	Class	Desirability score	Absolute ranking
CHEMBL4159660	active	0.98	1
CHEMBL4161976	active	0.96	2
CHEMBL3401707	inactive	0.96	3
CHEMBL4163037	active	0.96	4
CHEMBL3417645	active	0.95	5
CHEMBL4591649	active	0.94	6
CHEMBL4216419	active	0.94	7
CHEMBL4591060	active	0.90	8
CHEMBL4452963	active	0.90	9
CHEMBL4443887	active	0.88	10
CHEMBL3740457	active	0.86	11
CHEMBL3342758	active	0.85	12
CHEMBL4176257	active	0.83	13
CHEMBL1271386	active	0.82	14
Decoy	inactive	0.81	15
Decoy	inactive	0.81	16
Decoy	inactive	0.81	17
Decoy	inactive	0.80	18
Decoy	inactive	0.80	19
Decoy	inactive	0.80	20
Decoy	inactive	0.80	21
CHEMBL3264508	active	0.79	22

Conclusions

We have successfully constructed the protocol based on model assembling and desirability conversion to predict HepG2 cytotoxicity. Numerous base models were developed by random feature selection and SVM modelling. Base models yield excellent specificity and acceptable ROC, but their sensitivity is low of only around 0.30. Ten ensemble models proved to enhance the performance of base models in both training and test set referring to ROC by approximately 30 % and sensitivity by at least 60 %. By desirability transformation, 10 models ranked chemical compounds in validation dataset and exhibited great prediction power with early recognition (BEDROC in range of 0.54 to 0.63) and enrichment (EF in range of 25.00 to 29.55). The best protocol (model 1) has BEDROC of 0.63 and EF of 29.55 recalls 63.64 % of total known active compounds in the validation dataset in top 1 % fraction.

Acknowledgements

This study is supported by the National Foundation for Science and Technology Development of Vietnam (NAFOSTED) grant number 108.05-2017.13.

References

[1] A Koulouris, C Tsagkaris, V Spyrou, E Pappa, A Troullinou and M Nikolaou. Hepatocellular carcinoma: An overview of the changing landscape of treatment options. *J. Hepatocellular Carcinoma* 2021; **8**, 387-401.

[2] EN Muratov, J Bajorath, RP Sheridan, IV Tetko, D Filimonov, V Poroikov, TI Oprea, II Baskin, A Varnek, A Roitberg, O Isayev, S Curtalolo, D Fourches, Y Cohen, A Aspuru-Guzik, DA Winkler, D Agrafiotis, A Cherkasov and A Tropsha. QSAR without borders. *Chem. Soc. Rev.* 2020; **49**, 3525-64.

[3] M Cruz-Monteagudo, F Borges and MNDS Cordeiro. Desirability-based multiobjective optimization for global QSAR studies: Application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. *J. Comput. Chem.* 2008; **29**, 2445-59.

[4] CA Coello, SG Brambila, JF Gamboa, MGC Tapia and RH Gómez. Evolutionary multiobjective optimization: Open research areas and some challenges lying ahead. *Complex Intell. Syst.* 2020; **6**, 21-36.

[5] VV Kleandrova, MT Scotti, L Scotti, A Nayarisseri and A Speck-Planche. Cell-based multi-target QSAR model for design of virtual versatile inhibitors of liver cancer cell lines. *SAR QSAR Environ. Res.* 2020; **31**, 815-36.

[6] NM O'Boyle, M Banck, CA James, C Morley, T Vandermeersch and GR Hutchison. Open babel: An open chemical toolbox. *J. Cheminformatics* 2011; **3**, 33.

[7] D Castillo-González, JL Mergny, AD Rache, G Pérez-Machado, MA Cabrera-Pérez, O Nicolotti, A Introcaso, GF Mangiatordi, A Guédin, A Bourdoncle, T Garrigues, F Pallardó, MNDS Cordeiro, C Paz-y-Miño, E Tejera, F Borges and M Cruz-Monteagudo. Harmonization of QSAR best practices and molecular docking provides an efficient virtual screening tool for discovering new G-quadruplex ligands. *J. Chem. Inform. Model.* 2015; **55**, 2094-110.

[8] Y Perez-Castillo, A Sánchez-Rodríguez, E Tejera, M Cruz-Monteagudo, F Borges, MNDS Cordeiro, H Le-Thi-Thu and H Pham-The. A desirability-based multi objective approach for the virtual screening discovery of broad-spectrum anti-gastric cancer agents. *PLos One* 2018; **13**, e0192176.

[9] F Sahigara, K Mansouri, D Ballabio, A Mauri, V Consonni and R Todeschini. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 2012; **17**, 4791-810.

[10] R Polikar. Ensemble based systems in decision making. *IEEE Circ. Syst. Mag.* 2006; **6**, 21-45.

[11] MM Mysinger, M Carchia, JJ Irwin and BK Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* 2012; **55**, 6582-94.

[12] JF Truchon and CI Bayly. Evaluating virtual screening methods: Good and bad metrics for the "Early recognition" problem. *J. Chem. Inf. Model.* 2007; **47**, 488-508.

[13] S Kwon, H Bae, J Jo and S Yoon. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics* 2019; **20**, 521.

[14] P Pradeep, RJ Povinelli, S White and SJ Merrill. An ensemble model of QSAR tools for regulatory risk assessment. *J. Cheminform.* 2016; **8**, 48.