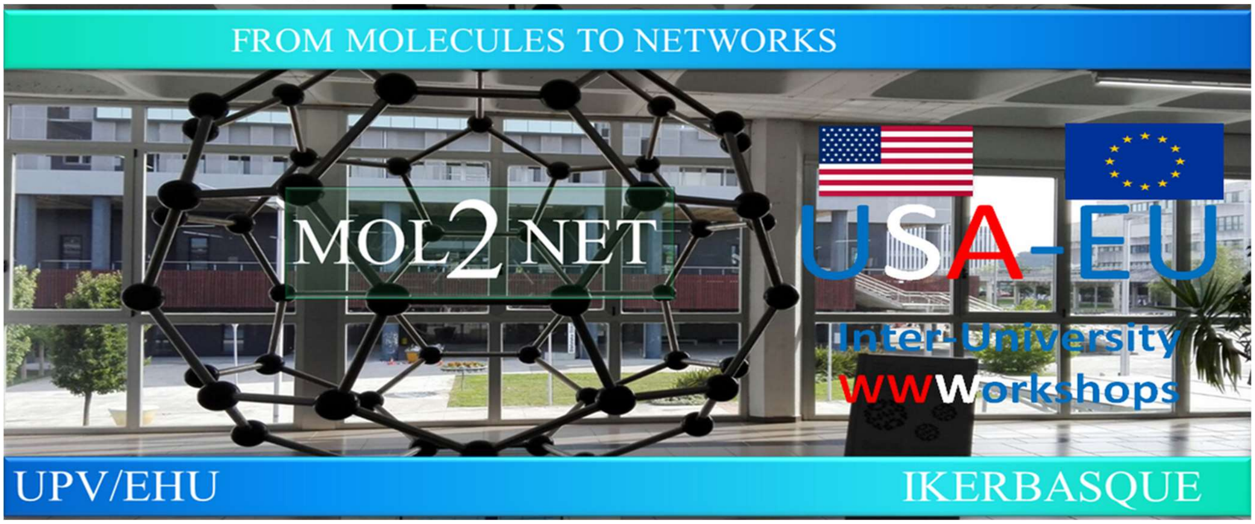
MOL2NET, International Conference Series on Multidisciplinary Sciences



Multiobjective screening of non-small cell lung cancer drug candidates

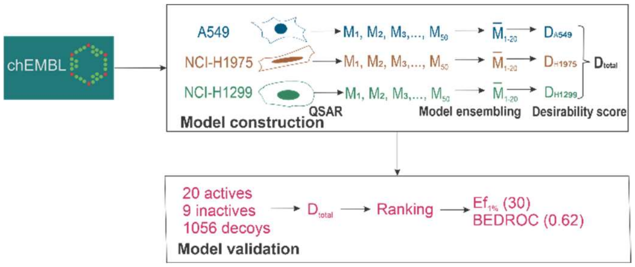
NGUYEN Phuong Nhung[a], *TRAN Dinh Nghia*[a], *LE Dinh Quang*[a], *NGUYEN Thi Kieu Oanh*[b], *PHAM The Hai*[c] .

[a] *General and Inorganic Department, Hanoi University of Pharmacy, 13 – 15 Le Thanh Tong Street, Hoan Kiem District, Hanoi, Vietnam*

[b] *Department of Life Science, University of Science and Technology of Hanoi, 18 Hoang Quoc Viet Road, Cau Giay District, Hanoi, Vietnam*

[c] *Medicinal Chemistry Department, , Hanoi University of Pharmacy, 13 – 15 Le Thanh Tong Street, Hoan Kiem District, Hanoi, Vietnam*

Graphical Abstract



Abstract.

Despite improvements in diagnosis and chemotherapy, non-small cell lung cancer (NSCLC) remains one of the most common cancer and has the largest proportion of all cancer death rates today. Computational approaches have been widely applied for early detection of novel treatment for NSCLC. Herein we developed a multi-objective approach or the screening of chemical compounds simultaneously active against three NSCLC cell lines: A549, NCI-H1299 and NCI-H1975. The first step consisted of developing ensemble models based on cytotoxicity data against three NSCLC cell lines curated from ChEMBL database. A desirable-based algorithm was then applied to incorporate these models into a multi-objective optimization system that can be used for virtual screening protocol. This system showed suitable screening performance with the Boltzmann-Enhanced Discrimination of ROC BEDROC = 0.62, the Enrichment Factor (EF)$_{1\%}$ = 30 and the Area Under the Accumulation Curve (AUAC) = 0.69

Introduction

With 2.1 million newly diagnosed cases and 1.8 million death in 2018, undoubtedly, lung cancer is the most common cancer type and leading cause of cancer death worldwide. 5 year survival rate of lung cancer is only 10 – 20% and varies significantly depending on the stage at diagnosis. The disease can be induced by many carcinogens including tobacco smoking asbestos, silica, several heavy metals, radon, and air pollution. Out of three histologic types of lung cancer, non-small cell lung cancer (NSCLC) accounts for majority of cases[1].

By exploitation of several targetable pathways such as EGFR, PI3K/AKT/mTOR, RAS–MAPK, and NTRK/ROS1 pathways, targeted therapy has been strongly developed. Some drugs based on this strategy are official approved and considered as first line treatment to replace traditional chemotherapy including gefitinib and erlotinib which are two first generations of EGFR – TKI (tyrorine kinase inhibitor of EGFR) and everolimus inhibiting of PI3K/AKT/mTOR pathway. Moreover, multiple chemical compounds and even microRNAs also show positive results in preclinical and clinical trials. Neverthless, similar to other pharmacotherapy, the cancer cell has adapted and resisted against these drugs. Although combination of different therapy has recommended[2], the effectiveness of current approach is still not completely raising the need of continuous effort of new anticancer drug.

On the other hand, multiple subtypes with a wide variety of morphological heterogeneity of lung cancer especially in lung adenocarcinoma[1] and the resistance of current targeted therapy[2] suggest that anticancer drug development should focus on multitarget rather than single targeted strategy. Multitarget drug development (MTDD) not only elevates the probabilities of interfering desired phenotype or achieving the therapeutic effects but also allows to optimize the toxicological profiles as well as bioavailability parameters. Furthermore, the complexity of pathological network of cancer cell brings it the ability to resist to drug therapy by compensation of signaling loops and makes the attempt of single node intervention ineffective[3]. In that scenario, MTDD by approaching simultaneously multiple nodes could be promising to knock down the cancer network and suppress the drug resistance.

In this study, our aim is to build a virtual model to reveal the small molecules with simultaneous activity against three non small cell lung cancer cell lines: NCI-H1299, NCI-H1975, and A549.

Materials and Methods

Dataset

All molecules for model construction of three cell line activity (NCI-H1299, NCI-H1975, and A549) were retrieved from ChEMBL database [4] and IC50 was used as activity indicator with threshold for classification of 10 µM. NCI-H1299, NCI-H1975, and A549 dataset have 520, 1131, and 21975 compounds respectively. Molecules were partitioned into training (70%), testing 1 (15%), and testing 2 (15%).

Base models

2D molecular descriptors were taken from CDK Descriptor Calculator. The number of selected descriptors of NCI – H1299, NCI-H1975, and A549 are 43, 46, and 35 respectively.

For multi-objective screening, firstly we developed base models to generalize the relationship between activity and structure for each cell line separately. In each model, we chose randomly 5, 10, and 15 descriptors, then constructed the model by RF algorithm[3]. We also defined the borderline for our models (applicability domain) by distance based method as per Mahalanobis distance[5].

Ensemble models

The reliability of prediction from single models could be significantly strengthen by integrating them into an ensemble model. In our study, we aggregated the base models by score vote strategy [3,6]. Outputs of active class of the models were averaged.

For model selection, we applied genetic algorithm (GA) approach with RF as learning method. The number of search iterations was 50 and at each iterations, 2 subsets were evaluated. The crossover and mutation probability was respectively 0.7 and 0.3. The performance of GA model was validated by 10 fold cross validation.

Desirability score

In the next step, we transformed the aggregated score to desirability score [3]. Desirability value of a compound (D_i) is defined as:

$$D_i = \begin{cases} 0 \text{ if } \widehat{S_\iota} \leq \min(S) \\ \left[\dfrac{\widehat{S_\iota} - \min(S)}{\max(S) - \min(S)}\right]^s \text{ if } \min(S) < \widehat{S_\iota} < \max(S) \\ 1 \text{ if } \widehat{S_\iota} \geq \max(S) \end{cases} \qquad (1)$$

which $\widehat{S_\iota}$ is the aggregated score of the new sample.

As our final objective is to conclude about the simultaneous activity against multiple cell lines, we continued aggregating the desirability of each end point into a final value, denoted as D_i^{total} by the following approach [3]:

$$D_i^{total} = \left[\prod_{j=1}^{n} D_{i,j}\right]^{\frac{1}{n}} \qquad (2)$$

Virtual screening of multi objective target

Molecules with known multi-effect on three cell lines were ranking by their total desirability scores. Decoys were generated by DUD-E database (http://dude.docking.org/generate) [7]. The efficiency of VS pipeline was evaluated by following parameters: Area Under the Accumulation Curve (AUAC), Enrichment Factor (EF), and Boltzmann-Enhanced Discrimination of ROC (BEDROC)[3].

Results and Discussion

Performance of base models is exhibited in *Figure 1*. For all cell lines, ROC of two testing sets was comparable except for the significantly lower ROC of testing set 2 of NCI-H1299. ROC of training dataset is higher than that of two testing datasets for NCI-H1299 and A549 but for NCI-H1975, two testing sets have better ROC values but the difference is not so high. Also, almost all models show good predictive power with ROC of higher than 0.65. ROC of NCI-H1299 and NCI-H1975 models is even better with values in range of 0.75 to 0.95. The sensitivity of models is very good and does not fluctuate substantially especially for NCI-H1975. Meanwhile, the specificity varies around 0.50 with some outliers of 0.1.

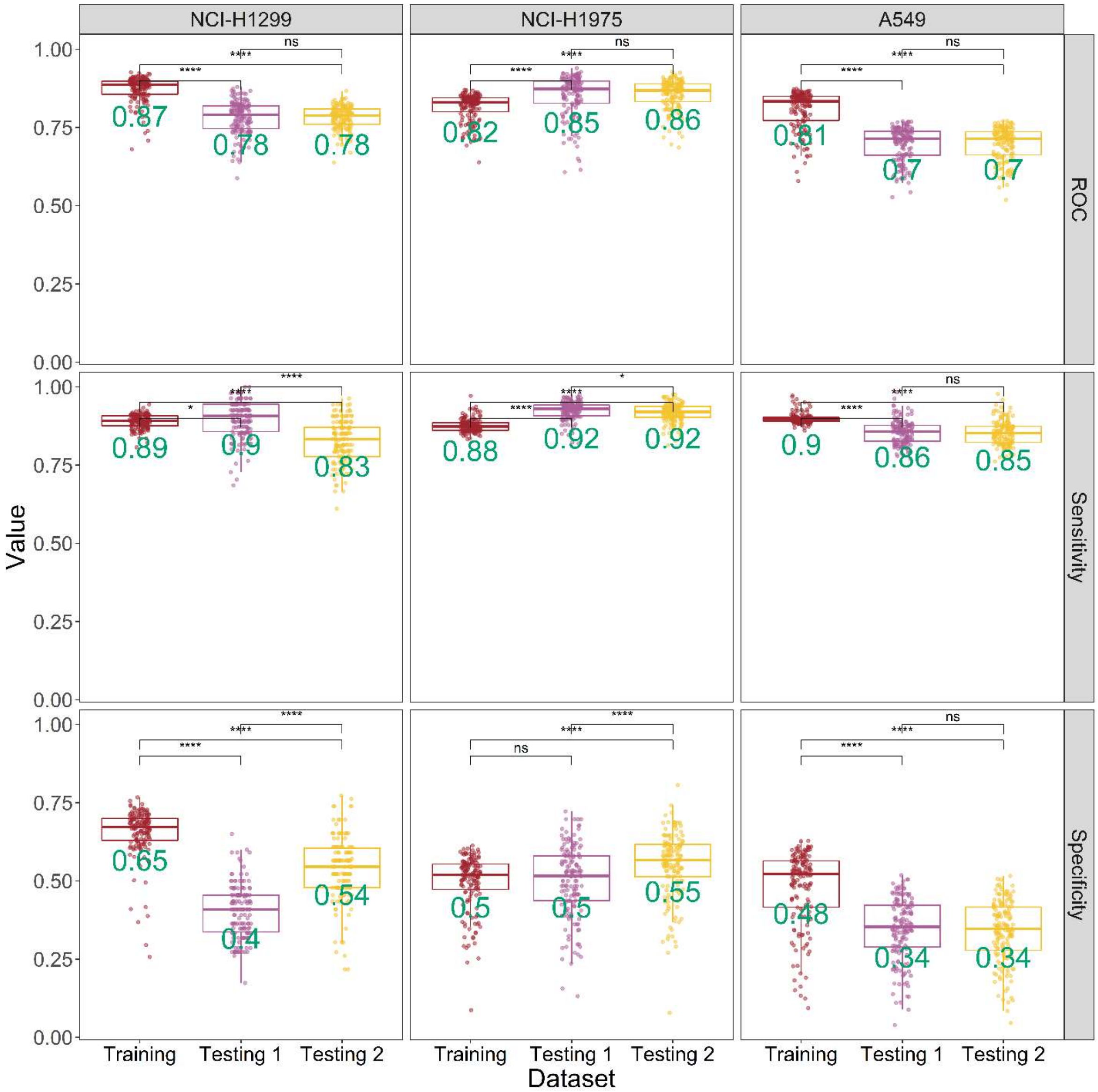


Figure 1. Performance of base models. Performance metrics of datasets were compared and the significance of difference was denoted by symbol (ns: not significant, *: significant with p value limit of 0.05, **: significant with p value limit of 0.01, ***: significant with p value limit of 0.001, ****: significant with p value limit of 0.0001).

For the first impression of the efficacy of model aggregation, we illustrated the relative improvement of esemble models on the basis of their base model performance mean in *Figure 2*.

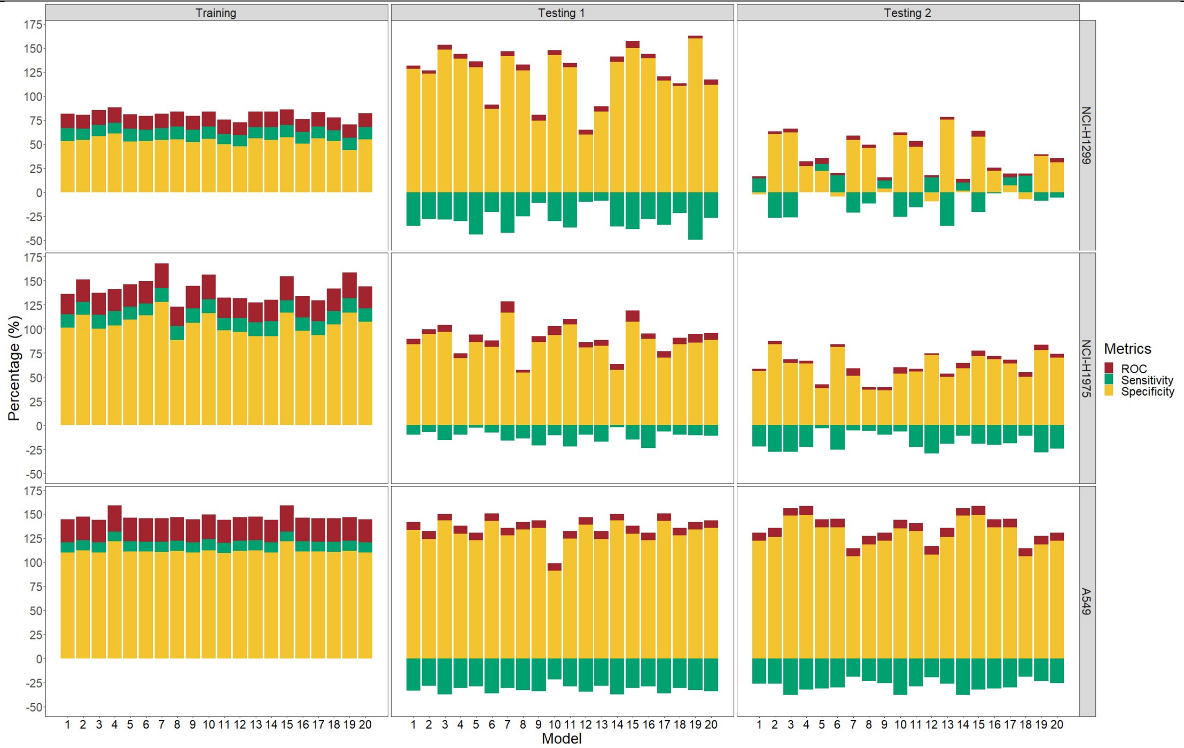


Figure 2. Relative improvement of ensemble models compared to their component means.

The outperformance of ensemble models is obvious. For the training set, all metrics are improved with the significant increase of specificity (around 100%). Although base models exhibit good ROC and sensitivity, these two parameters are still boosted with modest change of about 10 to 20%. The margin of improvement correlates with the size of the dataset. Meanwhile for two testing sets, ROC and specificity remain the same trend as the training set but sensitivity does not. ROC changes of testing sets are approximately half of that of training set but the enhancement of specificity is larger than training set for A549 and the testing 1 set of NCI-H1299 with some outstanding values of more than 150%. For NCI-H1975, specificity is also improved but with smaller margin. Sensitivity of ensemble models is poorer than that of base model means but the amount of decrease is only about 25% even in some cases (testing 2 of NCI-H1299), it does not decrease but shows constrast trend.

From 20 models per each endpoints, in total we generated 8000 possible combinations and the top VS models are shown in *Table 1*..

As the first target of virtual screening, we evaluated the ability of a VS protocol to maximize the total number of active compounds in a selected fraction of dataset through EF[3]. With the focus on the top 1% of screened data, all top models share the good EF with the maximum value of 30. Three remaining models have fairly lower EF of 25. This reflects the efficiency of our VS models to enrich 30 times more active molecules in the top 1% fraction than a uniform distribution of active compounds throughout the data.

Among four models, we further compared their power in ranking active compounds as close to the first position as posible which is called early recognition problem[3] by BEDROC. We assumed that 80% of screening importance is in the first 1% fraction corresponding to α value of 160.9. Again all models show remarkable performance with BEDROC of more than 0.60 and one model is outstanding with BEDROC of 0.62 (VS1).

Table 1. Virtual screening results

Protocol	AUAC	EF[1]	BEDROC[2]
VS1	0.69	30.00	0.62
VS2	0.73	30.00	0.60
VS3	0.71	30.00	0.60
VS4	0.67	30.00	0.60

[1]: EF at 1%

[2]: BEDROC with α = 160.9

The best BEDROC model successfully identifies active compounds in the first 6 positions. Although three models (VS2, VS3, VS4) yield the same BEDROC of 0.60, VS4 fails to recognize the sixth active position, VS2 and VS4 cannot identify the fifth ranking as active compound. However, if considering the highest ranking of inactive compounds, VS3 is the best model as inactive molecules are absent in the first 4.5% of screened data. Meanwhile with VS2, the first inactive compound is found in the first 2% fraction and with VS1 first 3% fraction. Balancing the efficiency of early recognition, we selected VS1 as the final VS model.

Conclusions

In conclusion, we developed a multi objective desirability based scheme for VS of NSCLC anticancer drug. The selected VS model (VS1) show outstanding performance referring two important targets of VS: maximizing the number of active compounds in the first 1% fraction of the data and early recoginition of active compounds.

References *(mandatory)*

1. Bernard, W. S. & Christopher, P. W. *World cancer report 2020. World Health Organization* (2020).
2. Yuan, M., Huang, L.-L., Chen, J.-H., Wu, J. & Xu, Q. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. *Signal Transduct. Target. Ther.* **4**, 61 (2019).
3. Perez-Castillo, Y. *et al.* A desirability-based multi objective approach for the virtual screening discovery of broad-spectrum anti-gastric cancer agents. *PLoS One* **13**, e0192176 (2018).
4. *CHEMBL database release 27.* http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_27 (2020) doi:10.6019/CHEMBL.database.27.
5. Sahigara, F. *et al.* Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **17**, 4791–4810 (2012).
6. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **6**, 21–45 (2006).
7. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).