Discovery of VEGFR-2 Inhibitors employing Junction Tree Variational Autoencoder with Bayesian Optimization and Gradient Ascent

Gia-Bao Truong,[†] Thanh-An Pham,[†] Van-Thinh To,[†] Hoang-Son Lai Le,[†] Phuoc-Chung Van Nguyen,[†] The-Chuong Trinh,[‡] Tieu-Long Phan,[¶,§] and Tuyen Ngoc Truong[*,†]

[†]*Faculty of Pharmacy, University of Medicine and Pharmacy at Ho Chi Minh City, 41 Dinh Tien Hoang, District 1, Ho Chi Minh City, 700000, Viet Nam*

[‡]*Faculty of Pharmacy, Grenoble Alpes University, La Tronche, FR 38700, France*

[¶]*Bioinformatics Group, Department of Computer Science & Interdisciplinary Center for Bioinformatics & School for Embedded and Composite Artificial Intelligence (SECAI), Leipzig University, Härtelstraße 16-18, Leipzig, DE 04107, Germany*

[§]*Department of Mathematics and Computer Science, University of Southern Denmark, Odense M, DK 5230, Denmark*

E-mail: truongtuyen@ump.edu.vn

Abstract

In the development of anticancer medications, the Vascular Endothelial Growth Factor Receptor 2 (VEGFR-2), which belongs to the protein tyrosine kinase family, emerges as one of the most significant targets of interest. The ongoing Food and Drug Administration (FDA) approval of novel therapeutic medicines towards VEGFR-2 emphasizes the urgent need to discover sophisticated molecular structures that are capable

of reliably limiting VEGFR-2 activity. Recognizing the huge potential of deep learning-based molecular model advancements, we focused our study on exploring the chemical space to find small molecules potentially inhibiting VEGFR-2. To achieve this goal, we utilized the Junction Tree Variational Autoencoder in combination with two optimization approaches on the latent space: the local Bayesian optimization on the initial dataset and the gradient ascent on nine FDA-approved drugs targeting VEGFR-2. The optimization results yielded a set of 493 uncharted small molecules. Quantitative structure-activity relationship (QSAR) models and molecular docking were used to assess the generated molecules for their inhibitory potential using their predicted pIC_{50} and binding affinity. The QSAR model constructed on RDK7 fingerprints using the CatBoost algorithm achieved remarkable coefficients of determination (R^2) of 0.792 ± 0.075 and 0.859 with respect to internal and external validation. Molecular docking was implemented using the 4ASD complex with optimistic retrospective control results (the ROC-AUC value being 0.710 and the binding activity threshold being -7.90 kcal/mol). Newly generated molecules possessing acceptable results corresponding to both assessments were shortlisted and checked for interactions with the protein at the binding site on important residues, including Cys919, Asp1046, and Glu885.

Keywords

Vascular endothelial growth factor receptor 2; junction tree variational autoencoder; bayesian optimization; gradient ascent

1 Introduction

According to GLOBOCAN 2022, cancer remains a predominant cause of mortality globally, accounting for approximately 10 million cases.[1] The incidence of cancer in the same year was reported to be 20 million cases. Considering static cancer rates, the number of new cases was estimated to reach over 35 million by the year 2050, which is a 77% increase from the

year 2022.

Tyrosine kinase receptors, particularly Vascular Endothelial Growth Factor Receptor 2 (VEGFR-2), represent a critical focal point in the development of anticancer drugs. VEGFR-2 plays a pivotal role in vasculogenesis and angiogenesis, processes co-opted by malignant tumors to facilitate the formation of new blood vessels, thereby enabling nutrient delivery and inducing metastasis.[2] VEGFR-2 inhibitors have demonstrated clinical efficacy across a broad spectrum of cancer types, including renal cell carcinomas, thyroid cancers, hepatocellular carcinomas, and soft tissue sarcomas, etc. as outlined in the NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines).[3–6] The ongoing trend of new therapeutics agents targeting VEGFR-2 approved by Food and Drug Administration (FDA) highlights the critical need to discover and develop advanced molecular entities capable of effectively inhibiting VEGFR-2 activity.[7]

Many *in silico* studies were made to identify compounds that have the potential to inhibit the VEGFR family using various computational tools.[8–14] However, nearly all these discoveries are virtual screenings from public or in-house databases. Furthermore, the quantitative structure-activity relationship (QSAR) models, molecular docking, and other techniques mainly acted as molecular filters in virtual screening. In contrast, regarding *de novo* approaches, these methods can help navigate the exploration in the chemical space to actively identify favorable compounds in a directed, properties-guided manner.[15,16]

Generative deep learning models are garnering significant attention in the discipline of *de novo* drug design due to their ability to extensively simulate and navigate the chemical space.[17,18] Capable of intricately reconstructing molecules from chemical space, these models play a crucial role in guiding the discovery process for new pharmaceutical candidates.

The chemical space is discrete and has no definite shape. Therefore, when Gómez-Bombarelli et al. created a molecular variational autoencoder (VAE) to optimize the molecular properties by which chemical structures are represented as a continuous vector space,[19] it paved the way for a series of VAE models applied to chemical spaces later.[20–22]

The Junction Tree Variational Autoencoder (JTVAE), pioneered by Jin et al.,[22] represents a significant breakthrough in molecular VAE models. This model innovatively segments the molecular graph into subgraphs, which subsequently become clusters on a junction tree. With this perspective, the JTVAE can decipher almost all meaningful simplified molecular-input line-entry system (SMILES) strings within the limit of its vocabulary set, surpassing previous capabilities of molecular VAEs. However, latent space of VAE is considered as a high-dimensional space. There are two common approaches to handle molecular optimization on this space: Bayesian optimization (BO) and gradient ascent (GA). BO aims to build a probabilistic surrogate model, usually a Gaussian process (GP),[23] to learn the distribution of the objective function on the latent variables. With the help of an acquisition function, BO on latent space iteratively samples and evaluates latent vectors in an efficient probability-based manner to generate optimal candidates.[19,24–26] On the other hand, the GA technique enables exploring the surrounding space of selected starting points in constrained optimization utilizing an auxiliary network.[19,22] In other words, BO helps exploit the favourable molecules regarding the extensive coverage of the latent space, while the GA method attempts to explore the promising specific chemical subspaces.

Acknowledging the promise of these innovations, our research focused on the exploration of the chemical space, with specific objective of identifying small molecules potentially inhibiting VEGFR-2. This was achieved through the application of local BO and GA methodologies to the JTVAE model. By combining the two approaches, we looked forward to achieving optimal molecules in the latent space and exploring the promising candidates derived from approved drugs on this target. The novel substances identified in our study were anticipated to exhibit key characteristics akin to those of approved VEGFR-2 inhibitors, yet with enhanced properties. To rigorously assess the potential of these newly identified molecules, we established a comprehensive QSAR model based on the methodologies developed by Phan et al. alongside a comprehensive molecular docking protocol.[27]

2 Results and discussion

2.1 Overview

In our study, we successfully employed the BO and GA techniques on the JTVAE model to explore the chemical space in order to generate batches of compounds for VEGFR-2 inhibitors' discovery. In particular, the BO approach was conducted with two objective functions, including pIC_{50} predicted using a QSAR model and *dual* value, which is a combination of predicted pIC_{50} and penalized logP (see Section 3.4), to yield 275 unique molecules. On the other hand, the GA algorithm resulted in 218 new substances from nine FDA-approved drugs as the starting points (see Section 3.5). A QSAR model was built using RDK7 fingerprints and the CatBoost (CatB) algorithm with good performance in predicting pIC_{50} values on VEGFR-2, which acted as a scoring function in BO tasks and assessed the potential of the generated molecules in the later steps. The compounds resulting from optimization underwent screening and potential ranking steps consisting of Molecular Sets (MOSES) filters,[28] QSAR, molecular docking, and *candidate* scoring to give 435 compounds potentially inhibiting VEGFR-2. The overview of these steps is illustrated in Figure 1.

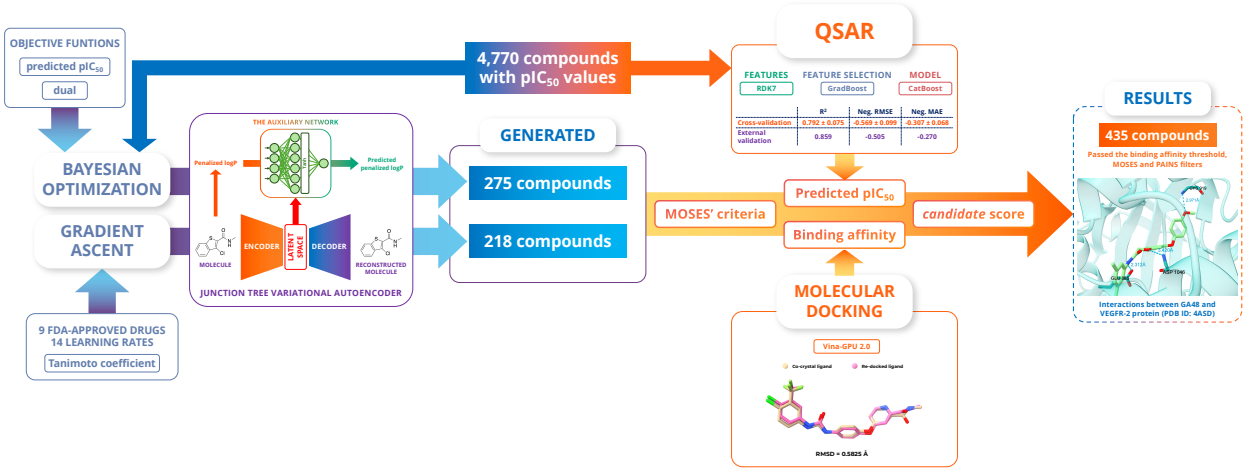


Figure 1: Overall workflow of discovering VEGFR-2 inhibitors using JTVAE utilizing BO and GA approaches. Compounds were generated from the two optimization tasks on the latent space of JTVAE. QSAR model, molecular docking and other criteria were used to assess the potential of these molecules in inhibiting VEGFR-2.

2.2 Pretraining of Variational Autoencoder

The JTVAE was trained and evaluated using MOSES datasets and criteria.[28] The MOSES training dataset yielded a vocabulary set of 533 SMILES strings representing molecular subgraphs. The evaluation of the pretraining model is summarized in Table 1. According to Table 1, the pretrained JTVAE model showed a strong performance across MOSES criteria, achieving near-perfect results. The model consistently generates valid and unique molecules. Notably, the novelty score is approximately equal to 1.0, indicating that the model is not overfitting. Furthermore, the reconstruction accuracy of the pretrained model, at 0.7048, is comparable to that of the JTVAE model in the original study by Jin et al., which reported a reconstruction accuracy of 0.767 using a different training dataset consisting of 250K molecules extracted from ZINC15 molecular database, compared to around 1.5M molecules in MOSES training dataset (see Section 3.1.1).[19,20,22,29]

Table 1: Pretrained JTVAE model evaluation.

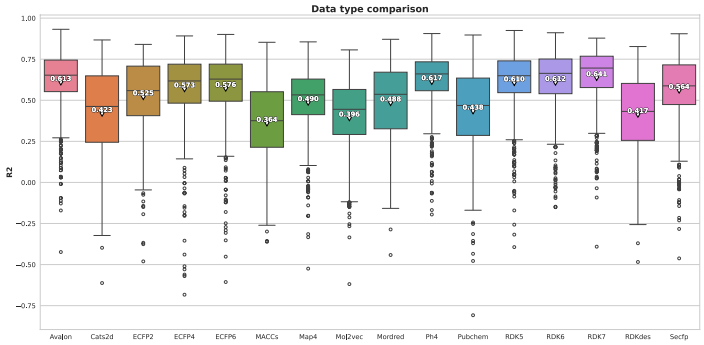
Criteria	**Value**
SMILES validity	1.0 ± 0.0
Unique SMILES in 1,000 samples	1.0 ± 0.0
Unique SMILES in 10,000 samples	0.9998 ± 0.0002
Novelty	0.9671 ± 0.0005
Reconstruction accuracy	0.7048

2.3 Quantitative structure-activity relationship model

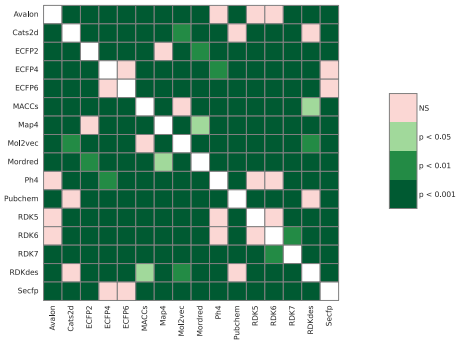
2.3.1 Feature engineering

The VEGFR-2 data was collected and preprocessed as described in Section 3.1.2. According to the *meta-analysis* result, the dataset using RDKit fingerprint (pathway fingerpring) with $maxPath = 7$ (RDK7) gave the highest mean of coefficient of determination (R^2) of 0.641 (Figure 2A), which was significantly higher than the others ($p < 0.01$) based on Wilcoxon signed-rank tests (Figure 2B).

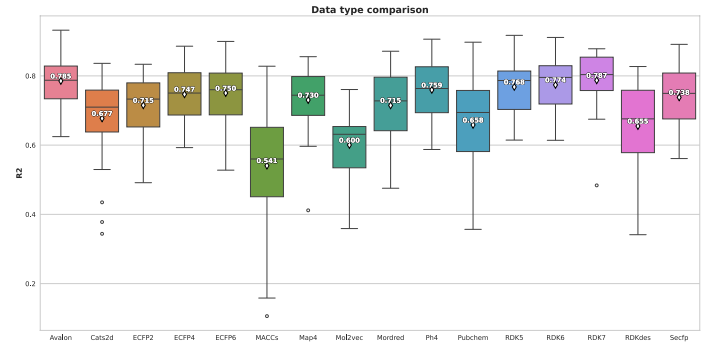
A. Boxplot comparing 16 types of features in meta-analysis (R^2).



B. Wilcoxon signed-rank tests of meta-analysis.



C. Boxplot comparing 16 types of features in subgroup analysis (R^2).



D. Wilcoxon signed-rank tests of subgroup analysis.

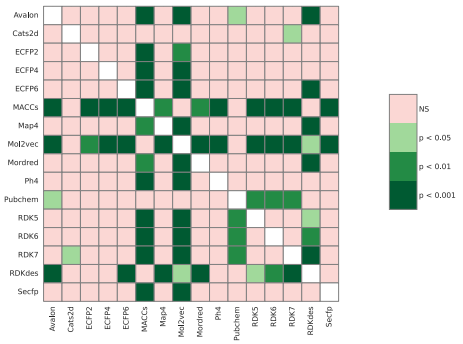


Figure 2: Analysis of 16 types of molecular representations based on R^2. **A.** Boxplot comparing 16 types of features in meta-analysis; **B.** Heatmap of the Wilcoxon signed-rank tests in meta-analysis; **C.** Boxplot comparing 16 types of features in subgroup analysis; **D.** Heatmap of the Wilcoxon signed-rank tests in subgroup analysis.

In addition, *subgroup analysis* comparing the performance of the best algorithms for each dataset, has also shown that the RDK7 dataset still produced a high average R^2 result of 0.787 (Figure 2C). While this result was the highest compared to the other molecular features, it only showed statistically significant differences when compared with the CAT2D,[30] MACCS,[31] MOL2VEC,[32] PUBCHEM,[33] and RDKDES datasets ($p < 0.05$) (Figure 2D). Following the *meta-analysis* and *subgroup analysis* results, RDK7 molecular fingerprints were selected for constructing the QSAR model.

We performed data duplication cleaning and applied a variance threshold of 0.05 to both datasets, yielding 4,110 data points in the training set while the external validation set remained unchanged. Feature selection was executed using the Gradient boosting (GRAD) algorithm because its performance was among the best in terms of the lowest number of features at 298 (Figure S1 and Table S2).

2.3.2 Model selection

The results of internal cross-validation of 15 algorithms (Figure 3) demonstrated that the CATB[34] algorithm emerged as the most proficient in terms of R^2. Particularly, CATB achieved an R^2 value of 0.792 ± 0.075, marking a statistically significant enhancement over its counterparts, as evidenced by the Wilcoxon signed-rank analysis ($p < 0.01$), illustrated in Figure 3B. Moreover, the CATB algorithm's effectiveness is further corroborated by negative root mean square error ($RMSE$) and negative mean absolute error (MAE) values, recorded at -0.569 ± 0.099 and -0.307 ± 0.068, respectively (Figure S2).

2.3.3 External validation

The outcomes of external validation mirrored those of the internal cross-evaluation, affirming the CATB algorithm as a robust predictive model, as evidenced by its R^2, $RMSE$, and MAE values of 0.859, 0.505, and 0.270 respectively. A comparative analysis of the five top-performing algorithms from internal cross-validation, including k-nearest neighbors

Figure 3: Internal cross-validation results of 15 machine learning models based on R^2. **A.** Boxplot comparing 15 machine learning models; **B.** Heatmap of the Wilcoxon signed-rank tests.

(KNN), Random Forest (RF), XGBoost (XGB) and Histogram gradient boosting (HIST),[35] demonstrated CATB's superior generalization capabilities, unequivocally conforming its superior performance, detailed in Table 2. Consequently, the CATB algorithm was selected for constructing QSAR model.

Table 2: External validation results of the five best performing algorithms. Better model has greater R^2 value, lower $RMSE$ value or lower MAE value.

Model	R^2 (↑)	$RMSE$ (↓)	MAE (↓)
KNN	0.849	0.523	**0.232**
RF	0.852	0.517	0.265
XGB	0.833	0.550	0.301
HIST	**0.859**	0.506	0.278
CATB	**0.859**	**0.505**	0.270

2.4 Bayesian optimization

Molecules with pIC_{50} values greater than or equal to 6.0 were extracted from the QSAR dataset to obtain the active initial set consisting of 1,360 substances. Moreover, a further refinement using a cutoff of pIC_{50} values greater than 8.0 yielded an alternative set of 1,127 substances. We conducted 11 experimental runs where each run yielded a final result of k substances identified as having the highest target function values. The number of best k substances and novel substances not previously present in the dataset across these runs is summarized in Table S4.

In our study, we identified 358 novel molecules that were not present in the initial datasets. After removing duplicates using SMILES strings, we isolated a diverse final set of 275 unique substances. The predicted pIC_{50} values of these substances ranged from 4.36 to 7.31, accounting for 55.0% of the total compounds analyzed across all experiments. These molecules underwent further evaluation through molecular docking to explore their efficacy in inhibiting VEGFR-2.

2.5 Gradient ascent

In comparison to BO method which required a sufficient size of the initial dataset together with objective function values, GA method only needs a small set of substances whose activities on the chosen target are confirmed, such as approved drugs, or lead compounds. The GA method was used on nine FDA-approved drugs as starting points to obtain 218 different molecules, whose predicted pIC_{50} values ranged from 4.50 to 7.40. A brief overview of the results is given in Figure 4 as two-dimensional Principal Component Analysis (PCA) visualization using RDK7 fingerprints. To emphasize the impact of the learning rates, the generated molecules from the sorafenib exploration were visualized in Figure 4, where molecules resulted from greater learning rates locate further to the initial point. The numbers of found molecules corresponding to each starting point are listed in Table S5.

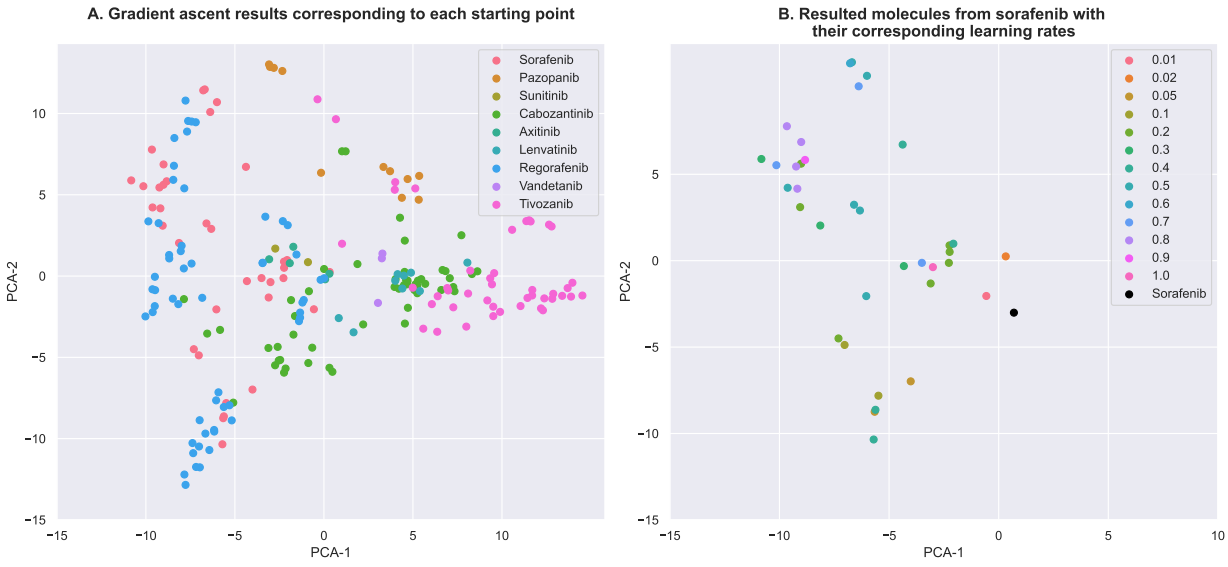


Figure 4: 2D PCA of the GA results. **A.** All resulted molecules corresponding to each starting point; **B.** All resulted molecules from the sorafenib exploration corresponding to their configured learning rates. PCA-1 and PCA-2 are the corresponding axes of the data points resulting from PCA dimensionality reduction algorithm.

However, compared to other staring points, the numbers of molecules resulting from sunitinib, axitinib, and vandetanib (8, 33, and 11 molecules before removing duplicates, respectively) were remarkably lower (Table S5). This may highlight the importance of placing

starting points in the latent space, where they may fall into uncertain or uneven subspaces. In other words, changes in these starting points' latent vectors could lead to significant structural changes in the generated molecules, affecting their Tanimoto similarities with respect to their starters. Furthermore, as the learning rate increases, the latent vectors undergo stronger changes, according to the algorithm (Algorithm 1). Therefore, we applied constraints on Tanimoto similarity, and only a limited number of learning rates were considered in this study (Section 3.5).

Notably, there was no overlap between the molecules generated via GA and those from BO, cumulatively resulting in a total of 493 novel molecules through these two optimization approaches. Subsequent molecular docking studies were conducted to evaluate their potential as inhibitors of VEGFR-2.

2.6 Molecular docking

To assess the efficacy of our docking model, we employed re-docking and enrichment analysis to evaluate its docking and screening capabilities, respectively. The re-docking results, depicted in Figure 5D, demonstrate that the re-docked ligand aligns closely with the co-crystal ligand, achieving a root mean square deviation ($RMSD$) of 0.5825 Å, which is considered excellent given that it is below the 2.0 Å threshold.

We utilized Butina clustering on the QSAR training dataset, employing active molecules targeting VEGFR-2, to identify 10 centroids, as depicted in Figure 5A. A scatter plot was made to represent these clusters by using t-Distributed Stochastic Neighbor Embedding (t-SNE)[36] on 50-dimensional vectors resulting from PCA on RDK7 fingeprints (Figure 5B), confirming their scaffolds' diversity. This diversity is crucial for minimizing analog bias in decoy generation. Using these centroids along with sorafenib, the active ligand in the protein complex with ID: 4ASD, we generated 550 decoys with an active-inactive ratio of 1:50. These were subsequently analyzed to determine the optimal binding affinity threshold via area under the receiver operating characteristic curve (ROC-AUC) comparison, shown in

Figure 5C. The *median* score type achieved the highest ROC-AUC value of 0.710, setting the binding affinity threshold at -7.90 kcal/mol.

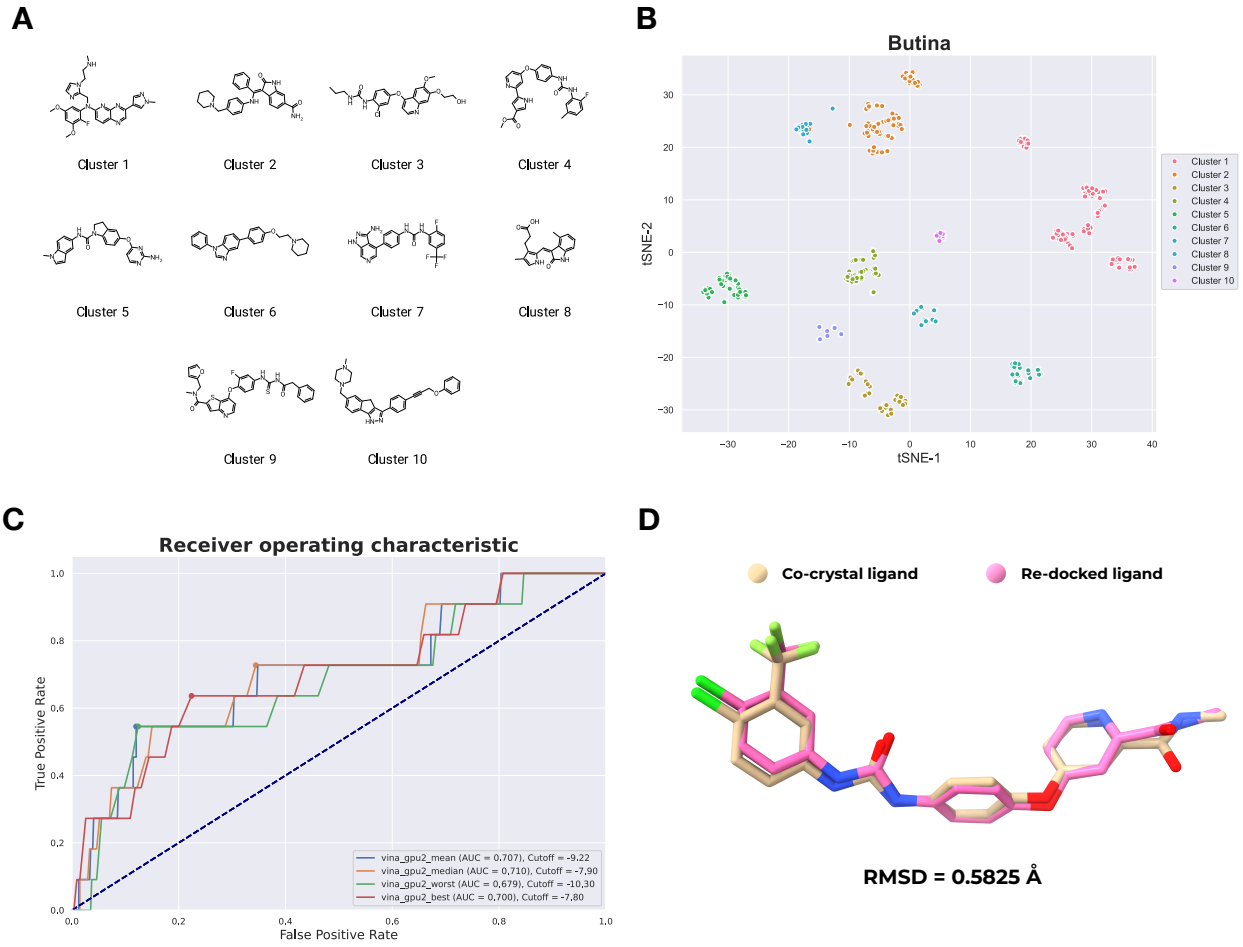


Figure 5: Butina clustering of QSAR training dataset. **A.** Centroids obtained from Butina clustering; **B.** Scatter plot of clusters. tSNE-1 and tSNE-2 are the corresponding axes of the data points resulting from t-SNE dimensionality reduction algorithm; **C.** Retrospective control internal validation results; **D.** Re-docking results.

2.7 Analysing generated molecules

To begin with, we applied the same MOSES filters on the resulted molecules from optimization experiments as in the data preparation step (Section 3.1.2) to remove molecules violating the MOSES criteria, especially the Pan Assay Interference Compounds (PAINS) filter,[37] culminating in a set of 435 molecules. The QSAR model and molecular docking

pipeline were applied to these molecules to calculate the predicted pIC_{50} and binding affinity. The *candidate* score values were calculated separately for each optimization method. Top 10 of these molecules corresponding to BO and GA results based on the *candidate* score are presented in Figure 6 along with their assessment results.

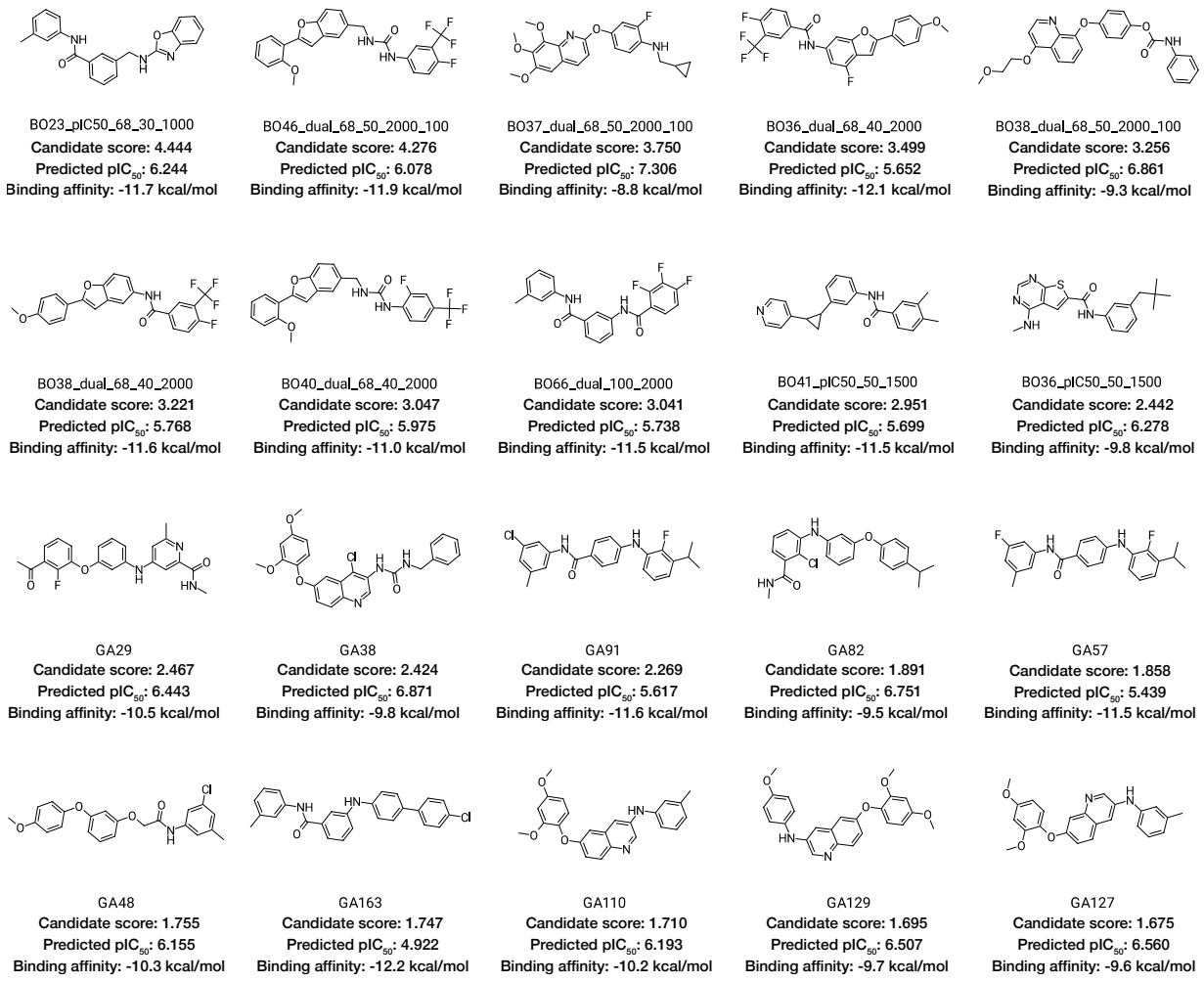


Figure 6: Top ranked molecules gathered from optimization experiments based on the *candidate* score. The first 10 molecules were obtained from BO, and the next 10 were from GA.

Inferred from the molecular docking results, most of these top-ranked molecules were found to form either hydrogen bonds or Van der Waals interactions with Cys919, Asp1046 or Glu885, which are crucial residues in inhibiting VEGFR-2 (Table S6). Notably, GA48, with a docking score of -10.3 kcal/mol and a predicted pIC_{50} of 6.155 (Figure 6), formed a hydrogen bond with the backbone of Cys919 through the oxygen atom of its methoxy group,

and hydrogen bonds with the side chain of Glu885 and Asp1046 (Figure 7). Similarity, BO23_pIC50_68_30_1000 and BO66_dual_100_2000 achieved promising docking scores of -11.7 kcal/mol and -11.5 kcal/mol, respectively. While they did not form hydrogen bonds with Cys919, interacting only through Van der Waals forces, they established new hydrogen bonds with the side chain of Cys1045 via the oxygen atom of their amide groups within the binding site. These interactions contribute to the rigidity of the potential compounds within the binding pocket. In terms of top-performing molecules in BO experiments, seven out of the top ten molecules came from runs that had their initial dataset cut off.

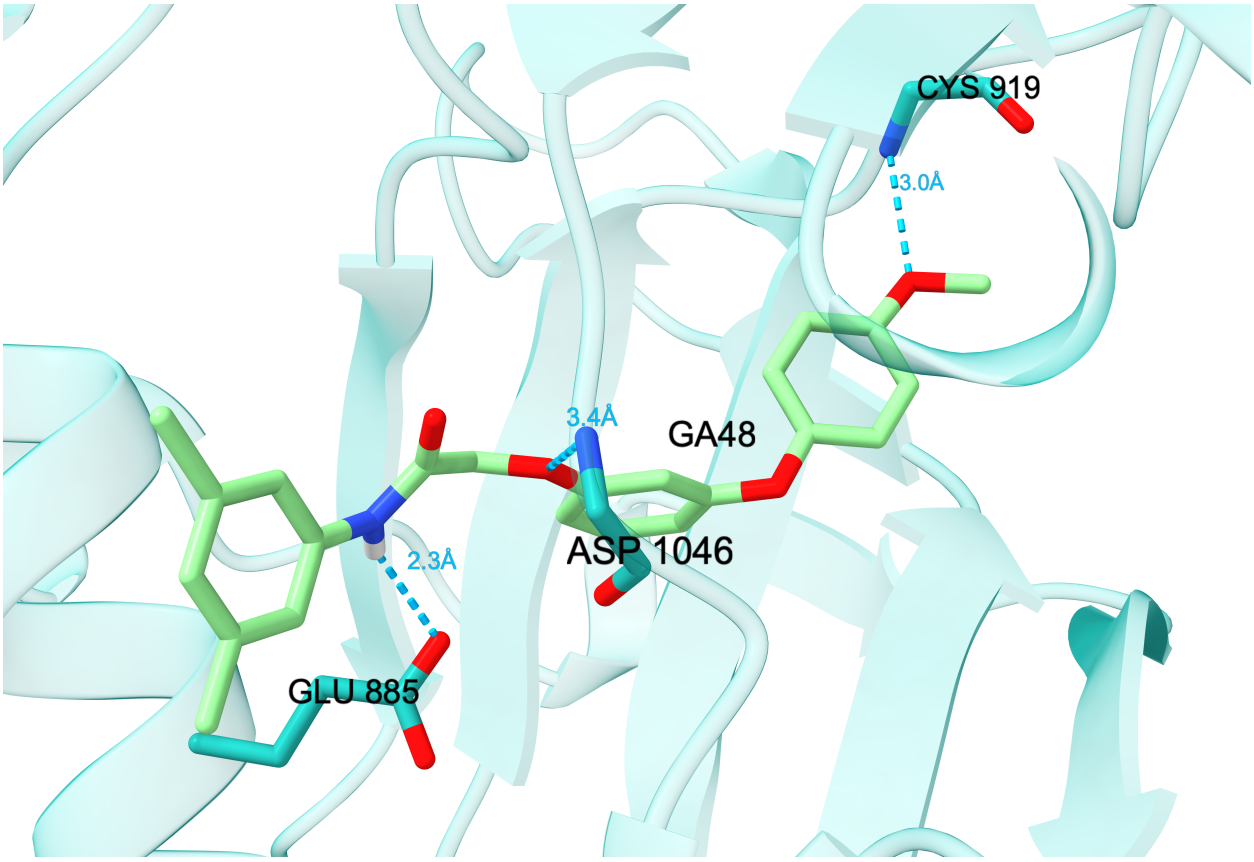


Figure 7: Molecular docking visualization for GA48. The GA48 molecule formed hydrogen bonds to all three crucial residues consisting of Cys919, Asp1046 and Glu885.

In further analysis, the differences between the two types of objective functions, the predicted pIC_{50} and the *dual* values, in BO were also taken into account. While the *dual* function values may not be limited in a range compared to the predicted pIC_{50}, it could drive

the optimization process to focus on generating molecules that are unfavourable hydrophobic, since the logP values are not upper bounded. For example, molecules from a certain run using *dual* values as the objective function are shown in Figure S3. These molecules, while syntactically and grammatically correct within the dataset, do not correspond to viable chemical structures. They have chains of phenyl rings, contain a low amount or even absence of either nitrogen, oxygen or halogen atoms, and are highly hydrophobic. This phenomenon could also happen to GA method when optimization is done without constraints such as molecular similarity. Fortunately, the problem could be avoided by utilizing the initial dataset for BO, experimenting with different configurations, applying constraints or other assessments that can discern and reject chemically infeasible configurations. Therefore, a pIC_{50} cut-off dataset and multiple settings were proposed for BO, and Tanimoto similarity was applied to GA. The Local latent space Bayesian optimization (LOLBO) method were also applied to help limiting the search space to prevent over-exploration of areas that are unreliable. The *candidate* score could also counter this phenomenon, since unfavourable molecules could have either low predicted pIC_{50} values or bad binding affinity. For instance, unfavourable molecules could have good binding affinity due to overflowed Van der Waals interactions, but low predicted pIC_{50} values, resulting in poor *candidate* scores (Figure S3).

3 Methods

3.1 Data

3.1.1 Variational autoencoder training

The dataset employed for VAE was sourced from the MOSES[28] dataset, which is comprised of molecules curated and screened from the ZINC20 database.[38] The training set encompassed a substantial collection of 1,584,663 substances from MOSES.

For the VAE evaluation, the validation dataset was inherited from the MOSES evalua-

tion datasets consisting of Test set (176,074 molecules) and TestSF set (176,225 molecules). The substances in these datasets were represented as canonical SMILES strings, processed using the `RDKit` library[39] without stereochemical information.

3.1.2 Molecular data targeting VEGFR-2

The dataset containing active substances was extracted from research articles and was available at `https://github.com/buchijw/QSAR-VEGFR2`. The final goal of filtering the dataset was to obtain substances which showed their *in vitro* toxicity or inhibitory activity toward the human VEGFR-2 strains. The activity types were chosen to be IC_{50} or pIC_{50}, then being standardized to the pIC_{50} values. Molecular SMILES strings were canonicalized and filtered using the `RDKit` library according to the MOSES data preparation criteria,[28,39] including uncharged atoms, constraints on certain allowed elements, number of atoms in cycles and medicinal chemistry filters, as well as containing no information about stereochemistry. Then, molecules also went through the JTVAE vocabulary check to ensure their encoding and decoding availability. Lastly, outlier removal was performed on the dataset using $z - score$ to obtain the final training dataset for the QSAR model.

3.1.3 Molecular docking material

The structure of human ligand-binding human VEGFR-2 protein was derived from the RCSB Protein Data Bank (`https://www.rcsb.org/`) with an identification code of 4ASD (PDB ID: 4ASD) (Figure 8).[40] This protein was used to perform molecular docking for substances generated in two techniques.

3.2 Junction Tree Variational Autoencoder

JTVAE is an unsupervised generative model consisting of an encoder $q_\phi(\mathbf{Z}|\mathbf{D})$ that maps a set of n discrete molecules $\mathbf{D} = \{x_i\}_{i=1}^n$ to a continuous latent space $\mathbf{Z}$ and a decoder $p_\theta(\mathbf{D}|\mathbf{Z})$ that does the reverse mapping. The parameters of the encoder ϕ and the decoder θ

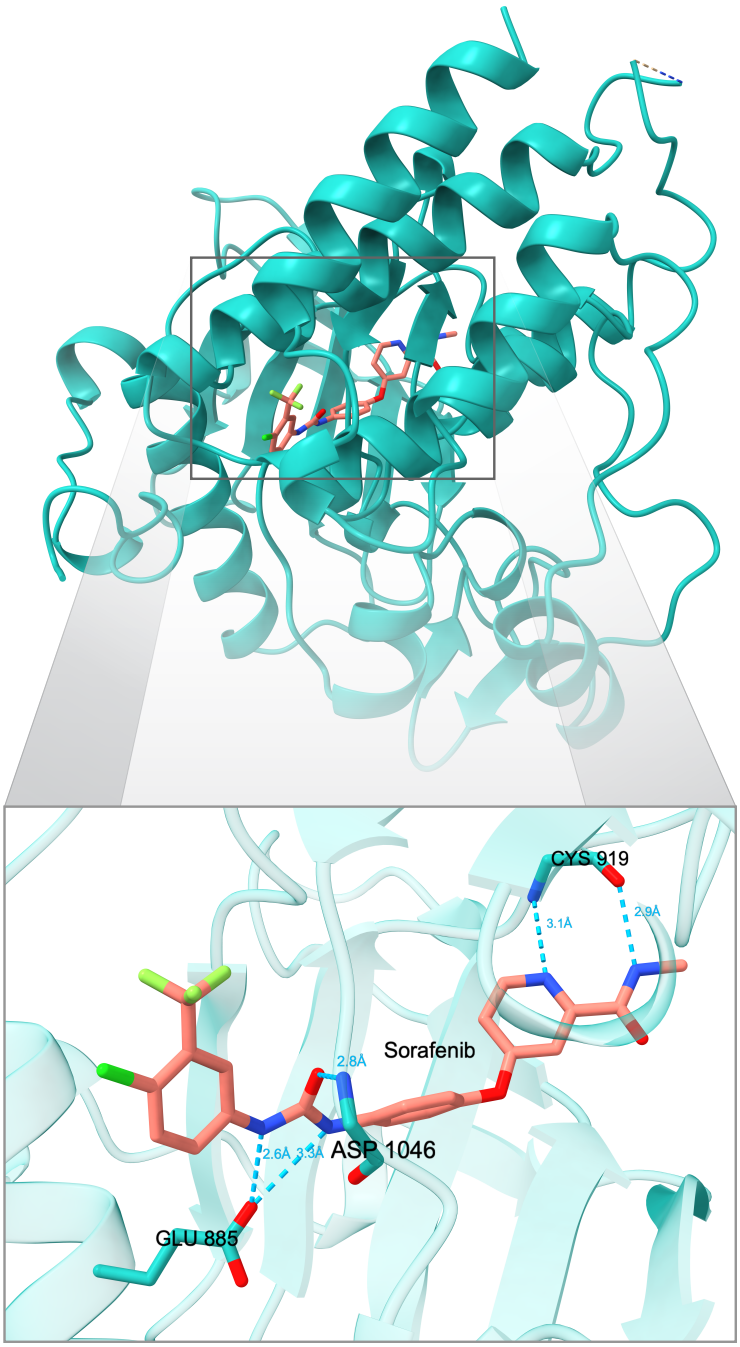


Figure 8: The 4ASD complex and the binding site of VEGFR-2 and sorafenib. The sorafenib ligand forms four hydrogen bonds with Cys919, Glu885 and Asp1046 residues of VEGFR-2.

are updated using the loss function $\mathcal{L}_{VAE}$. Compared to older molecular generative models, the JTVAE achieved the ability to generate almost all meaningful SMILES strings in the limit of its vocabulary set, due to exploiting molecular subgraphs as clusters on junction trees.

The JTVAE was implemented with a neural network that predicts the molecules' properties, as stated in the original paper,[22] and contributes as a critical component for facilitating constrained optimization using the GA technique (Figure 9). The auxiliary network incorporates a forward structure, initiated by a latent vector serving as the input. This is followed by a linear layer that outputs a 450-dimensional vector. Subsequently, a tanh activation layer processes this output, culminating in a predicted property value at the network's output. The output value is then compared with the original properties value to compute the loss function of the model based on mean squared error (MSE).

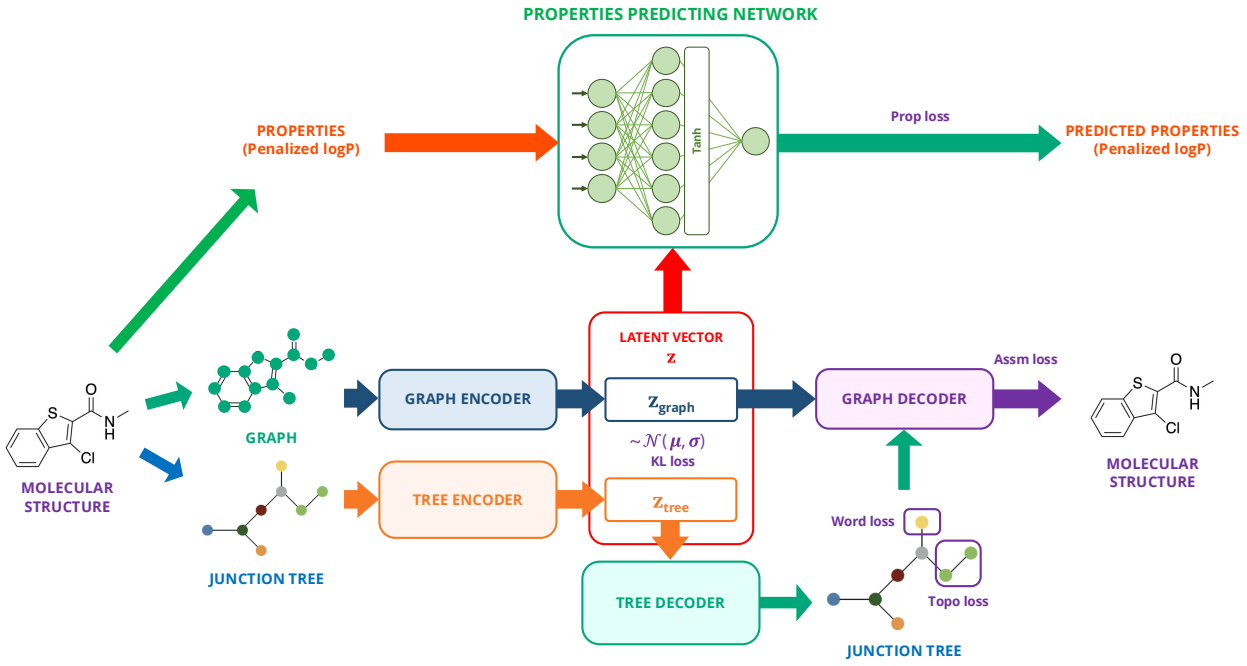


Figure 9: The general framework of the JTVAE. This diagram illustrates the comprehensive architecture of the JTVAE, detailing its neural network structure that predicts molecular properties. `Word loss`: Loss function of predicting clusters' labels; `Topo loss`: Loss function of predicting junction tree topology; `Assm loss`: Loss function of assembling molecular subgraphs; `KL loss`: Kullback-Leibler (KL) divergence loss function; `Prop loss`: Loss function of predicting molecular properties.

The training dataset was employed to establish the vocabulary set, encompassing the subgraph structures presented throughout the dataset. The standardized penalized logP value was chosen as the target molecular property. Penalized logP is a calculated molecular property expressing drug-likeness of a molecule,[20] which is a combination of octanol-water partition coefficients (logP), synthetic accessibility score (SA),[41] and cycle score (number of rings that have more than six atoms) that can be derived for a molecule m as:

$$penlogP(m) = logP(m) - SA(m) - cycle(m) \tag{1}$$

where $penlogP(m)$ is the penalized logP score for the molecule m, $logP(m)$, $SA(m)$ and $cycle(m)$ are the standardized values of its logP, SA score and cycle score, respectively.

On maximizing this value, one wants molecules with high logP for drug-likeness, low SA score and cycle score for viable synthesizability. In earlier studies, the goal for optimizing drug-likeness is frequently selected as the penalized logP value.[19–22]

The JTVAE training process was executed with a configuration featuring a hidden size of 450 and a latent size of 56, continuing until convergence was achieved. The training process encompassed 30 epochs, during which approximately 700,000 steps of model parameter updates were conducted. All configurations of this pretraining model are detailed in Table S3. Subsequently, model performance was validated using MOSES' established evaluation criteria, with 100,000 random molecular samples decoded from the latent space. To ensure robustness, this sampling and evaluation process was repeated three times, each with a different seed.

3.3 Quantitative structure-activity relationship model

The obtained VEGFR-2 dataset (Section 3.1.2) was utilized to train the QSAR model that predicts molecular pIC_{50} values (Figure 10). The substances in the dataset were conducted fingerprints and descriptors calculations from SMILES strings using the `RDKit` library[39] as

the input of the QSAR model. Here, we examined 16 types of molecular features including molecular descriptors and fingerprints. Each dataset was built and cross-evaluated using 15 different machine learning regression algorithms (Table S1). The *random_state* or *seed* variable was set to 42 to ensure reproducibility.

Figure 10: QSAR pipeline. At the beginning, features computed for the processed data went through duplication handling and feature cleaning in the data engineering stage. Next, in the feature engineering stage, *Meta-analysis* and *subgroup analysis* were employed to select the well-performed feature type. Then, feature selection was conducted to optimize the number of features. Analyses were made in the model selection step to choose the best performing algorithm. Lastly, the built QSAR model was external validated to assess the performance.

Molecules in the datasets were filtered using the JTVAE vocabulary set to ensure validity within the defined chemical space of the VAE model. We partitioned the data using a stratified scaffold splitting strategy, where molecules between training and external validation perceived non-overlapping Bemis-Murcko scaffolds while retaining the distributions of the target values in the two datasets.[42] With this strategy, splitting the data into an exact partition could be impossible. Therefore, we optimized the proportion of molecules in the test set to a range of 10-20%, which corresponds to ratios of training and test sets between 90:10 and 80:20. Particularly, we divided the dataset, which consisted of 4,770 substances represented by 16 types of molecular features, into a training set of 4,115 substances and an external validation set of 655 substances, corresponding to proportions of 86% and 14%, respectively. In the data engineering stage, features with low variance (less than 0.05) or those missing in over 50% of data were excluded; otherwise, they were imputed using KNN imputation with $k = 5$.

In order to determine the most suitable type of molecular features, the prepared dataset for each feature type went through a full QSAR pipeline with default configurations in the feature engineering stage. In this contribution, we define the term *meta-analysis* as the process where Wilcoxon statistical tests were employed to evaluate and compare the

performance across all algorithms for each dataset. Similarly, *subgroup analysis* is defined as the application of Wilcoxon statistical tests to assess and compare the performance among the top-performing algorithms for each dataset. Details of these processes are shown in Figure 11.

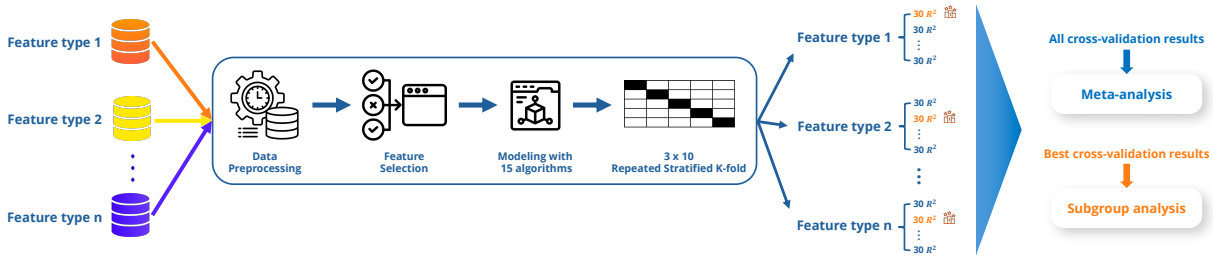


Figure 11: *Meta-analysis* and *subgroup analysis* processes.

Utilizing Wilcoxon statistical tests within the frameworks of *meta-analysis* and *subgroup analysis*, we selected the preferred feature type if it was statistically significantly better than the others in terms of R^2. Subsequently, feature selection or feature importance based dimensionality reduction was performed on seven algorithms including: analysis of variance (ANOVA), Mutual information (MUTUAL_INFO), RF,[43] Extremely Randomized Trees (EXT),[44] Adaptive boosting (ADA),[45] GRAD[46] and XGB[47] to reasonably reduce the number of features. The refined training dataset was underwent evaluation through internal 10-fold cross-validation, repeated three times with the stratified scaffold splitting strategy, to select the most effective machine learning algorithm. Both feature and model selection processes were done using Wilcoxon statistical tests based on R^2, $RMSE$ or MAE. Finally, external validation was performed on the external validation dataset using the chosen settings and three metrics for generalization assertment.

The final QSAR model was then implemented in the BO pipeline for the objective function calculation as well as obtaining the predicted pIC_{50} of the candidates for analysis.

3.4 Local Bayesian optimization

BO is well-known as a optimization method suitable for searching in a high dimensional space with objective functions that are expensive to calculate. In general, this technique aims to build a surrogate model $f(\mathbf{x})$ to approximate the distribution of the target function. This model is usually a GP $\mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, {}^{23}\mathbf{x}'))$ that is iteratively refined with new samples $\hat{\mathbf{x}}$ and their corresponding target function values $\hat{\mathbf{y}}$ using the Bayes' rule. Beside the surrogate model, an acquisition function is used to select the next candidates $\hat{\mathbf{x}}$ to observe that can bring the best updates to the surrogate model.

BO in molecular latent space requires inputs that are latent vectors $\mathbf{z}$ of a set of molecules after passing through the encoder of the VAE and their objective function values $\mathbf{y}$ to train the surrogate model $f(\mathbf{z})$. Then, the acquisition function suggests new latent vector candidates $\hat{\mathbf{z}}$ that are decoded back to molecules so their objective function values $\hat{\mathbf{y}}$ can be calculated. These new data points are then used to update the surrogate model $f(\mathbf{z})$ in the next iteration.

Compared to global BO methods, locally searching in the latent space can be more efficient since it can avoid over-exploration of areas that are unreliable. The trust region Bayesian optimization (TuRBO) algorithm,[26] which is at the heart of the LOLBO algorithm,[25] can use hyper-rectangular zones (trust regions) that are iteratively shrinking to narrow down the search areas and get closer to the optimal data points. As an advancement, LOLBO can navigate the latent space by jointly training the VAE with the surrogate model after multiple failed optimization attempts, enabling more effective exploration of better molecules.

In order to explore molecules based on their inhibitory activities on VEGFR-2, the objective function of the BO process was chosen to be whether the predicted pIC_{50} value resulted from the QSAR model, or the *dual* value calculated using the formula:

$$dual = pIC_{50_{norm}} + penlogP_{norm} \tag{2}$$

where $pIC_{50_{norm}}$ and $penlogP_{norm}$ are respectively the normalized values of pIC_{50} and penalized logP of the substances.

LOLBO was conducted using the trained JTVAE model with Sparse Gaussian Process (SparseGP) as the surrogate model and Thompson sampler.[48] We utilized the `lolbo` package to conduct BO.

For the initial set of molecules, we used substances whose pIC_{50} values were greater than or equal to 6.0 in the active dataset in order to provide the starting set of molecules with high activities on VEGFR-2. Moreover, due to the limitation of the QSAR model, we also tested configurations of cutting off substances with pIC_{50} values more than 8.0 to obtain an alternative initial set. The reason was that the QSAR model can not predict pIC_{50} values better than the highest one in the active dataset, therefore cutting off may encourage the algorithm to find novel molecules. The optimization was conducted 11 times with different configurations (Table S4). New substances found were then carried out molecular docking.

3.5 Gradient ascent

For constrained optimization, one optimization method is GA, which needs training a VAE along with an extra neural network that learns to predict the target function value $\mathbf{y}$ from the latent vector $\mathbf{z}$. The new latent vector $\mathbf{z}'$ can be calculated using the gradient of the target function $\mathbf{y}$ with respect to the latent vector $\mathbf{z}$ and the configurable learning rate α as in the equation (3). Then, new molecular candidates can be decoded from these modified latent vectors using the VAE decoder and their target function values can be calculated. These steps can be repeated for a certain number of iterations or until a condition is met.

$$\mathbf{z}' = \mathbf{z} + \alpha \times \nabla_{\mathbf{z}}\mathbf{y} \tag{3}$$

GA technique (1) was used to conduct constrained optimization on nine FDA-approved VEGFR-2 inhibitors as starting points including axitinib (Inlyta), cabozantinib (Cometriq),

lenvatinib (Lenvima), pazopanib (Votrient), regorafenib (Stivarga), sorafenib (Nexavar), sunitinib (Sutent), tivozanib (Fotivda), and vandetanib (Zactima). The target property was chosen to be the standardized penalized logP of the substances. We ran the exploration process with 14 different learning rates (α) ranging from 0.01 to 1.50 for each starting point. Each run consisted of 50 search iterations (n_iter) and only candidates whose Tanimoto similarities corresponding to their starting point passed the threshold were selected. The Tanimoto similarity threshold (sim_cutoff) was set to 0.3 and was calculated using extended-connectivity fingerprints with diameter 4 (ECFP4) fingerprints.[49]

Algorithm 1 Gradient ascent algorithm

Require: Starting molecule x, learning rate α, search iterations for each run n_iter, Tanimoto similarity threshold sim_cutoff

procedure GradientAscent($\alpha, n\_iter, sim\_cutoff$)
  $\mathbf{z_0} \leftarrow$ latent vector of x using VAE's encoder
  $fp_0 \leftarrow$ ECFP4 fingerprints of x
  Explored candidates $\mathcal{C} \leftarrow \emptyset$
  for $i \leftarrow 1, n\_iter$ **do**
    $\hat{y} \leftarrow$ predicted target property of $\mathbf{z_{i-1}}$ using VAE's auxiliary network
    $\nabla_{\mathbf{z_{i-1}}} \hat{y} \leftarrow$ gradient of $\hat{y}$ with respect to $\mathbf{z_{i-1}}$
    $\mathbf{z_i} \leftarrow \mathbf{z_{i-1}} + \alpha \times \nabla_{\mathbf{z_{i-1}}} \hat{y}$ ▷ Equation (3)
    $x_i \leftarrow$ decoded molecule of $\mathbf{z_i}$ using VAE's decoder
    $fp_i \leftarrow$ ECFP4 fingerprints of x_i
    $\mathcal{T} \leftarrow$ Tanimoto similarity between fp_0 and fp_i
    if $\mathcal{T} \geq sim\_cutoff$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{x_i\}$
    end if
  end for
  return $\mathcal{C}$
end procedure

3.6 Molecular docking

New substances found in the BO and GA processes were conducted molecular docking using `Vina-GPU 2.0` software[50] as shown in Figure 12. The VEGFR-2 - sorafenib complex (PDB ID: 4ASD)[40] and ligands was prepared using `PyMOL`[51] version 3.0.0, `MGLTools` [52] version 1.5.7 with Patch 1 and `RDKit`[39] package. In parallel, the process of ligand preparation commenced

with the conversion of compounds from SMILES format into two-dimensional structures using the *MolFromSmiles* function from `RDKit`. Subsequently, these 2D structures were transformed into three-dimensional conformations via the *EmbedMolecule* module, utilizing a consistent random seed of 42. Energy optimization of the ligands was achieved using the Merck Molecular Force Field (MMFF94) via the *MMFFOptimizeMolecule* function, with the process limited to a maximum of 10,000 iterations. The optimized ligands were then saved in .pdb format. Additional preparation steps, employing `MGLTools`, included the addition of hydrogens and Gasteiger charges to the ligands.

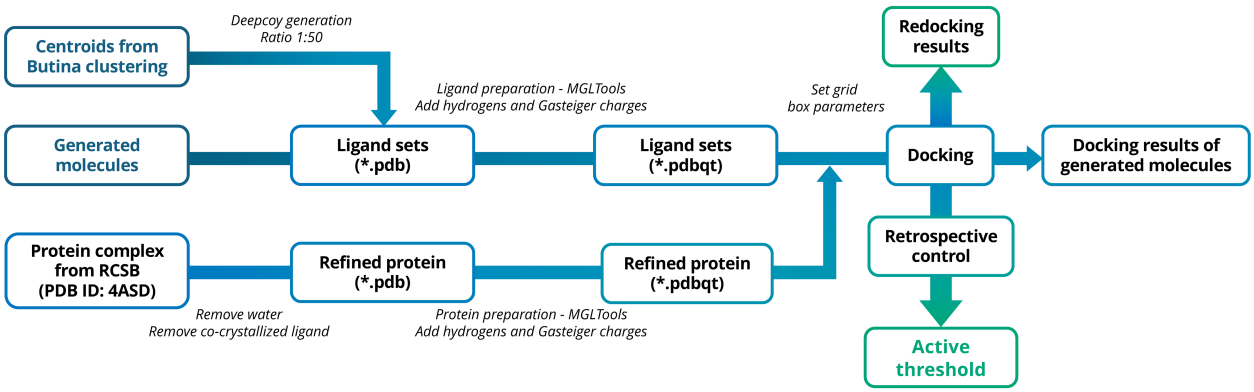


Figure 12: Molecular docking pipeline. `DeepCoy` was utilized for decoys generation from selected active molecules resulting from BUTINA clustering. Ligand sets including all actives, decoys and generated compounds were prepared by adding hydrogens and Gasteiger charges using `MGLTools`. Correspondingly, protein in the 4ASD complex was extracted and refined by separating from the co-crystallized ligand, water removing, and hydrogens and gasteiger charges addition. Finally, molecular docking was employed using the prepared ligand sets and protein. The results of redocking and retrospective control were used to assess the docking performance while the generated molecules' results were further analysed.

Docking grid box was configured to overlap Cys919, Glu885 and Asp1046 with the spacing of 0.375 Å. The grid box size was set to 17.9855 Å, 17.5803 Å, and 18.9919 Å, and the centers' coordinates were set to -23.2158 Å, 1.2313 Å, and -11.7709 Å, with respect to the X, Y, and Z axes. Other parameters included *seed* of 42, *thread* of 8000, *search_depth* of 50, and *num_modes* of 10. Binding affinities and analyses of interaction between ligands and the protein were used to assess the potential of the newly identified molecules.

To establish a binding affinity threshold for selection of promising generated molecules,

we initially applied the BUTINA clustering algorithm[53] to our QSAR training dataset to identify representative centroids of active molecules targeting VEGFR-2. Molecules whose pIC_{50} values were greater than or equal to 6.0 were considered active. Centroids together with sorafenib (molecule in the 4ASD complex) were then used to create decoys using `DeepCoy`[54] with an active-inactive ratio of 1:50 (Figure 12). Active molecules along with decoys were then docked using `Vina-GPU 2.0` software to obtain binding affinity cutoffs. The optimal threshold was determined to be the binding affinity cutoff of the highest geometric mean (g-mean) value corresponding to a certain scoring type. Generated molecules with binding affinity values better than the threshold were selected.

In order to rank generated molecules, we proposed a new scoring type called *candidate* score to combine the results from QSAR and molecular docking experiments. The *candidate* score was calculated as the sum of z-scores of the predicted pIC_{50} and the negative binding affinity values of the generated molecules. Then, the resulted compounds were ranked by the *candidate* score, in which molecules with higher *candidate* scores implied better potential on inhibiting VEGFR-2. Top ranked molecules were selected for further analysis.

4 Conclusions

In this study, we successfully identified several novel molecules by employing advanced techniques to explore and exploit chemical space. Despite this success, the vast complexity and sheer scale of chemical space present significant challenges. Our models, though generally effective, sometimes struggled with the diversity and complexity found in molecular databases. The accuracy and reliability of our predictions, especially concerning novel or underrepresented molecules, continue to be major concerns. Moreover, scalability is a critical issue; increasing the size and diversity of our datasets tends to exponentially increase computational demands. To overcome these challenges and enhance the accuracy of our exploratory models, it is essential to enrich the diversity of the training datasets. Such improvements

will not only boost the models' ability to generalize across various regions of chemical space but will also enhance their predictive accuracy.

5 Conflicts of interest

There are no conflicts to declare.

6 Data and Software Availability

The code and the data supporting this article can be found at `https://github.com/buchijw/JTVAE-lolbo-VEGFR2` for Bayesian optimization, `https://github.com/buchijw/JTVAE-GA` for gradient ascent and `https://github.com/buchijw/QSAR-VEGFR2` for QSAR. The `MGLTools` application can be found at `https://ccsb.scripps.edu/mgltools/`. The version of the `MGLTools` application employed for this study is version 1.5.7 with Patch 1. The `PyMOL` application can be found at `https://github.com/schrodinger/pymol-open-source`. The version of the `PyMOL` employed for this study is version 3.0.0. The `Vina-GPU 2.0` can be found at `https://github.com/DeltaGroupNJUPT/Vina-GPU-2.0`.

Acknowledgement

This work has received support from the Korea International Cooperation Agency (KOICA) under the project entitled "Education and Research Capacity Building Project at University of Medicine and Pharmacy at Ho Chi Minh City," conducted from 2024 to 2025 (Project No. 2021-00020-3). We also thank Ngoc-Vi Nguyen Tran for commenting on the manuscript.

Supporting Information Available

A listing of the contents of each file supplied as Supporting Information should be included. For instructions on what should be included in the Supporting Information as well as how to prepare this material for publications, refer to the journal's Instructions for Authors.

The following files are available free of charge.

- SI.pdf: Additional experimental configurations and results (PDF)
- Optimization_results_sorted.xlsx: Resulting molecules from Bayesian optimization and gradient ascent runs with their corresponding SMILES entries, predicted pIC_{50} values, docking scores and *candidate* scores (XLSX)

References

(1) Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca-Cancer J. Clin.* **2024**, *74*, 229–263.

(2) Olsson, A.-K.; Dimberg, A.; Kreuger, J.; Claesson-Welsh, L. VEGF receptor signalling ? in control of vascular function. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 359–371.

(3) NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Kidney Cancer (Version 3.2024). 2024; `https://www.nccn.org/professionals/physician_gls/pdf/kidney.pdf`. Accessed 1 May, 2024.

(4) NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Thyroid Carcinoma (Version 2.2024). 2024; `https://www.nccn.org/professionals/physician_gls/pdf/thyroid.pdf`. Accessed 1 May, 2024.

(5) NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Hepatocel-

lular Carcinoma (Version 1.2024). 2024; https://www.nccn.org/professionals/physician_gls/pdf/hcc.pdf. Accessed 1 May, 2024.

(6) NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Soft Tissue Sarcoma (Version 1.2024). 2024; https://www.nccn.org/professionals/physician_gls/pdf/sarcoma.pdf. Accessed 1 May, 2024.

(7) Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2024 update. *Pharmacol. Res.* **2024**, *200*, 107059.

(8) Al-Sanea, M. M.; Chilingaryan, G.; Abelyan, N.; Sargsyan, A.; Hovhannisyan, S.; Gasparyan, H.; Gevorgyan, S.; Albogami, S.; Ghoneim, M. M.; Farag, A. K.; Mohamed, A. A. B.; El-Damasy, A. K. Identification of Novel Potential VEGFR-2 Inhibitors Using a Combination of Computational Methods for Drug Discovery. *Life (Basel, Switz.)* **2021**, *11*.

(9) Yucel, M. A.; Adal, E.; Aktekin, M. B.; Hepokur, C.; Gambacorta, N.; Nicolotti, O.; Algul, O. From Deep Learning to the Discovery of Promising VEGFR-2 Inhibitors. *ChemMedChem* **2024**, *19*, e202400108.

(10) Salimi, A.; Lim, J. H.; Jang, J. H.; Lee, J. Y. The use of machine learning modeling, virtual screening, molecular docking, and molecular dynamics simulations to identify potential VEGFR2 kinase inhibitors. *Sci. Rep.* **2022**, *12*, 18825.

(11) Kang, D.; Pang, X.; Lian, W.; Xu, L.; Wang, J.; Jia, H.; Zhang, B.; Liu, A.-L.; Du, G.-H. Discovery of VEGFR2 inhibitors by integrating naïve Bayesian classification, molecular docking and drug screening approaches. *RSC Adv.* **2018**, *8*, 5286–5297.

(12) Alamri, M. A.; Merae Alshahrani, M.; Alawam, A. S.; Paria, S.; Kumar Sen, K.; Banerjee, S.; Saha, S. Development of newer generation Vascular endothelial growth factor Receptor-2 Inhibitors: Pharmacophore based design, virtual Screening, molecular

Docking, molecular dynamic Simulation, and DFT analyses. *J. King Saud Univ., Sci.* **2024**, *36*, 103285.

(13) Baammi, S.; El Allali, A.; Daoud, R. Unleashing Natures potential: a computational approach to discovering novel VEGFR-2 inhibitors from African natural compound using virtual screening, ADMET analysis, molecular dynamics, and MMPBSA calculations. *Front. Mol. Biosci.* **2023**, *10*.

(14) Fouad, M. A.; Osman, A. A.; Abdelhamid, N. M.; Rashad, M. W.; Nabawy, A. Y.; El Kerdawy, A. M. Discovery of dual kinase inhibitors targeting VEGFR2 and FAK: structure-based pharmacophore modeling, virtual screening, and molecular docking studies. *BMC Chem.* **2024**, *18*, 29.

(15) Meyers, J.; Fabian, B.; Brown, N. *De novo* molecular design and generative models. *Drug Discovery Today* **2021**, *26*, 2707–2715.

(16) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, e1608.

(17) Zeng, X.; Wang, F.; Luo, Y.; gu Kang, S.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **2022**, *3*, 100794.

(18) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular designa review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849.

(19) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(20) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. Proceedings of the 34th International Conference on Machine Learning. 2017; pp 1945–1954.

(21) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. 6th International Conference on Learning Representations. 2018.

(22) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. Proceedings of the 35th International Conference on Machine Learning. 2018; pp 2323–2332.

(23) Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*; Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 63–71.

(24) Notin, P.; Hernández-Lobato, J. M.; Gal, Y. Improving black-box optimization in VAE latent space using decoder uncertainty. Advances in Neural Information Processing Systems. 2021; pp 802–814.

(25) Maus, N.; Jones, H.; Moore, J.; Kusner, M. J.; Bradshaw, J.; Gardner, J. Local Latent Space Bayesian Optimization over Structured Inputs. Advances in Neural Information Processing Systems. 2022; pp 34505–34518.

(26) Eriksson, D.; Pearce, M.; Gardner, J.; Turner, R. D.; Poloczek, M. Scalable Global Optimization via Local Bayesian Optimization. Advances in Neural Information Processing Systems. 2019.

(27) Phan, T.-L.; Trinh, T.-C.; To, V.-T.; Pham, T.-A.; Van Nguyen, P.-C.; Phan, T.-M.; Truong, T. N. Novel machine learning approach toward classification model of HIV-1 integrase inhibitors. *RSC Adv.* **2024**, *14*, 14506–14513.

(28) Polykovskiy, D. et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*.

(29) Sterling, T.; Irwin, J. J. ZINC 15 Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337, PMID: 26479676.

(30) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.

(31) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(32) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35.

(33) National Institutes of Health (NIH) PubChem Fingerprint Description. 2024; `https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf`. Accessed 1 May, 2024.

(34) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems. 2018.

(35) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. 2017.

(36) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(37) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay

Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.

(38) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073.

(39) Landrum, G. et al. RDKit: Open-source cheminformatics. 2024.

(40) McTigue, M.; Murray, B. W.; Chen, J. H.; Deng, Y.-L.; Solowiej, J.; Kania, R. S. Molecular conformations, interactions, and properties associated with drug efficiency and clinical performance among VEGFR TK inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 18281–18289.

(41) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(42) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(43) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(44) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42.

(45) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.

(46) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232.

(47) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

(48) Titsias, M. Variational Learning of Inducing Variables in Sparse Gaussian Processes. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009; pp 567–574.

(49) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(50) Ding, J.; Tang, S.; Mei, Z.; Wang, L.; Huang, Q.; Hu, H.; Ling, M.; Wu, J. Vina-GPU 2.0: Further Accelerating AutoDock Vina and Its Derivatives with Graphics Processing Units. *J. Chem. Inf. Model.* **2023**, *63*, 1982–1998.

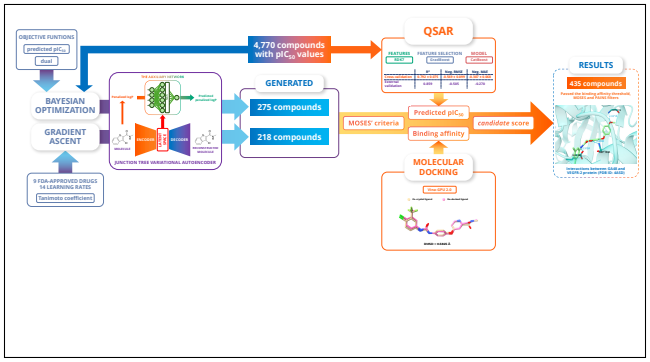
(51) Schrödinger, LLC The PyMOL Molecular Graphics System, Version 3.0. 2024.

(52) Center for Computational Structural Biology (CCRB) MGLTools. 2022; `https://ccsb.scripps.edu/mgltools/`.

(53) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.

(54) Imrie, F.; Bradley, A. R.; Deane, C. M. Generating property-matched decoy molecules using deep learning. *Bioinformatics* **2021**, *37*, 2134–2141.

TOC Graphic


4,770 compounds with pIC50 values
QSAR
BAYESIAN OPTIMIZATION
GRADIENT ASCENT
GENERATED
275 compounds
218 compounds
MOSES' criteria
Predicted pIC50
Binding affinity
candidate score
MOLECULAR DOCKING
RESULTS
435 compounds