Pharmaceutical Sciences and Drug Design
ISSN: 3062-4428
2025, Volume 5, Page No: 315-334
Copyright CC BY-NC-SA 4.0
Available online at: www.galaxypub.co/page/journals

Galaxy Publication

SIRT2i_Predictor: A Machine Learning-Driven Approach for Accelerating the Discovery of Selective SIRT2 Inhibitors in Age-Related Disease Therapeutics

Yasmin Rahman[1], Farid Ahmed[2]*, Shafiq Islam[1]

[1]Department of Pharmaceutical Sciences, School of Pharmacy, University of Dhaka, Dhaka, Bangladesh.
[2]Department of Drug Design, Faculty of Pharmacy, University of Malaya, Kuala Lumpur, Malaysia.

***E-mail** ✉ farid.ahmed@gmail.com

Received: 23 February 2025; **Revised:** 17 May 2025; **Accepted:** 24 May 2025

ABSTRACT

Recent preclinical findings have identified selective inhibitors of sirtuin 2 (SIRT2) as potential therapeutic agents for treating age-related diseases, but none have advanced to clinical trials. The growing adoption of machine learning (ML) techniques in drug discovery has demonstrated their transformative potential, yet there remains a lack of large-scale, robust ML models for identifying novel SIRT2 inhibitors. To fill this gap, we developed SIRT2i_Predictor, a machine-learning-based tool designed to assist in virtual screening (VS), lead optimization, and the selection of SIRT2 inhibitors for experimental validation. The tool integrates a series of high-performance ML models, both for regression and classification, to predict the potency of inhibitors and their selectivity across SIRT1-3 isoforms. These models were trained on an extensive dataset comprising 1797 compounds using state-of-the-art ML algorithms. A comparison with traditional structure-based VS protocols revealed that the tool not only covers a comparable chemical space but also offers significant improvements in processing speed. The tool was successfully applied to screen an in-house compound database, confirming its utility in prioritizing candidates for costly in vitro testing. With a user-friendly web interface**,** SIRT2i_Predictor is accessible to the broader research community, and its source code is freely available online.

Keywords: Machine learning, QSAR, Regression, Classification, Virtual screening, SIRT2 inhibitors

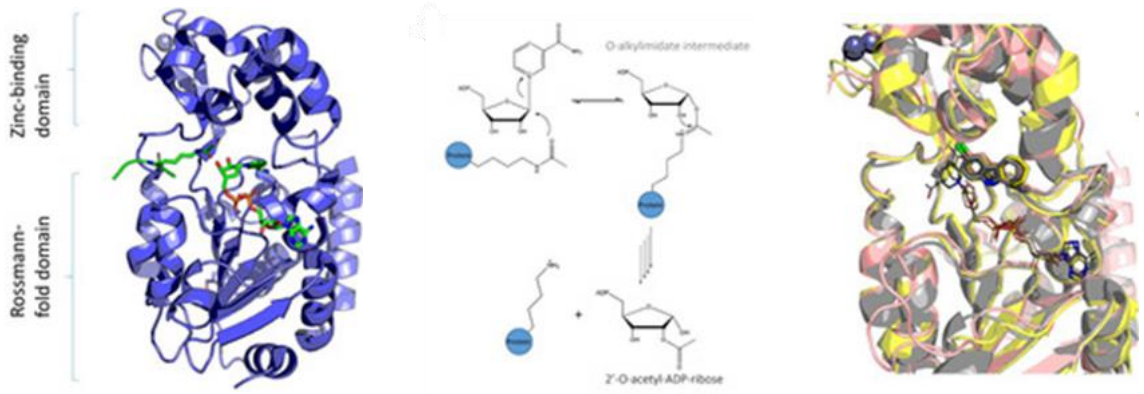
How to Cite This Article: Rahman Y, Ahmed F, Islam S. SIRT2i_Predictor: A Machine Learning-Driven Approach for Accelerating the Discovery of Selective SIRT2 Inhibitors in Age-Related Disease Therapeutics. Pharm Sci Drug Des. 2025;5:315-34. https://doi.org/10.51847/5ZwEspcuFd

Introduction

SIRT2 is an NAD+-dependent protein deacetylase crucial for regulating numerous biological processes, including genome stability, metabolism, aging, tumor development, and cell-cycle control [1-5]. Research using cellular and animal models has suggested that inhibiting SIRT2 could be a promising approach for treating age-related diseases, such as neurodegenerative conditions and cancer [6, 7]. Over the past ten years, growing preclinical evidence has increased interest in developing small-molecule inhibitors targeting SIRT2, with a particular focus on their potential as anticancer therapies [7]. Inhibition of SIRT2 has been found to play a key role in tackling various aspects of cancer progression, including inhibiting proliferation, invasion, angiogenesis, and metastasis [8-10]. Additionally, SIRT2 has been implicated in contributing to drug resistance in cancer therapy. Recent studies have shown that combining SIRT2 inhibitors with existing drugs such as dasatinib, doxorubicin, or paclitaxel could help overcome resistance in melanoma or certain subtypes of breast cancer cells [11-13]. Moreover, selective SIRT2 inhibitors have been explored as a means to enhance tumor immunotherapy by activating tumor-infiltrating lymphocytes, offering new possibilities for improving the clinical outcomes of TIL (tumor-infiltrating lymphocyte) and CAR-T (chimeric antigen receptor T-cell) therapies [14]. Despite decades of research and the discovery of many SIRT2 inhibitors, none have progressed to clinical trials, highlighting the need for further advancements in the field [15]. The most common challenges faced by existing inhibitors are poor selectivity, inadequate potency, or undesirable physicochemical properties [10, 15, 16].

© 2025 Pharmaceutical Sciences and Drug Design

Sirtuins, including SIRT2, have a catalytic core composed of a larger Rossmann-fold domain and a smaller zinc-binding domain, which are linked by flexible loops **(Figure 1a)**. All sirtuins operate through a shared catalytic mechanism, wherein a positively charged O-alkylimidate intermediate forms between NAD+ and the acetyl–lysine substrate, eventually leading to the hydrolysis of this intermediate into deacetylated polypeptides and 2′-O-acetyl-ADP-ribose **(Figure 1b)** [17]. The majority of known inhibitors target this catalytic mechanism by binding to the active site located in the cleft between the two domains **(Figure 1c).** However, due to the conserved nature of the catalytic site across sirtuin isoforms, achieving selectivity for SIRT2 inhibitors remains a significant challenge in drug development **(Figure 1c)** [15, 18]. Recent studies have pointed to the pharmacological benefits of selectively inhibiting SIRT2 over inhibiting other sirtuin family members, such as SIRT1 and SIRT3, which underscores the importance of selectivity in developing new SIRT2 inhibitors [19]. Additionally, the complex conformational flexibility of SIRT2 when interacting with inhibitors has been identified as a major barrier to discovering new compounds using structure-based computer-aided drug design (CADD) approaches [20]. Nevertheless, the extensive datasets generated in the search for novel SIRT2 inhibitors offer valuable opportunities for ligand-based CADD methods, particularly those incorporating machine learning techniques.



a) b) c)

Figure 1. Sirtuin Structure and Catalytic Mechanism Summary. (a) The two domains of sirtuins, demonstrated with the SIRT3 structure (PDB ID: 4FVT). NAD+ and the substrate are shown in green sticks. (b) The general mechanism of sirtuin-catalyzed deacetylation. (c) The challenge of achieving selectivity in sirtuin inhibitors, with aligned structures of SIRT1 (yellow) (PDB ID: 4I5I), SIRT2 (pink) (PDB ID: 5D7P), and SIRT3 (gray) (PDB ID: 4BV3). Some regions are omitted for clarity. Inhibitors (represented by gray, pink, or yellow sticks) display a consistent binding mode across all isoforms. NAD+ and ADP–ribose are shown in corresponding colors.

The discovery of novel therapeutic agents, particularly under the framework of precision medicine initiatives (NIH), increasingly relies on the integration of extensive datasets into drug development through cheminformatics [21]. The big data era in drug discovery has positioned artificial intelligence (AI) as a game-changing tool, capable of reducing both the time and costs associated with preclinical drug research [22, 23]. The rapid growth of machine learning (ML) applications in drug discovery has been facilitated by the availability of expansive datasets and the democratization of AI tools. Publicly available pharmacological databases, which continue to expand with records of biological activities, have allowed for more comprehensive approaches to drug discovery through ML-based modeling of structure–activity relationships (SAR) [21-23].

Quantitative structure–activity relationship (QSAR) modeling is a widely used computational method that correlates the structural properties of compounds with their biological activities, either through classification or regression models [24]. QSAR modeling has proven to be an effective strategy in preclinical drug development, particularly in predicting inactive compounds or minimizing side effects [21, 22, 25]. Analyses show that updating QSAR models with new data typically leads to better prediction accuracy and expanded applicability. QSAR models built on larger, more diverse datasets are able to cover a broader chemical space and offer more generalizable predictions. Consequently, large-scale QSAR models, trained on comprehensive datasets, are gaining popularity for their broad applicability and predictive power [25-27]. However, there remains a lack of robust and large-scale QSAR models for predicting the potency and selectivity of SIRT2 inhibitors. Developing

such models could significantly aid virtual screening, lead optimization, and drug repurposing efforts, as well as contribute to integrating cheminformatics with omics data in the increasingly complex precision medicine pipelines.

Given the promising preclinical data regarding the pharmacological benefits of selective SIRT2 inhibitors for various cancers, and their potential to enhance existing therapies such as immunotherapies, SIRT2 inhibitors could serve as a valuable addition to the arsenal of precision medicine drugs. To support the development of these inhibitors, this study focuses on creating a framework for efficient screening and evaluation of new compounds for their ability to inhibit SIRT2. This framework, called SIRT2i_Predictor, is based on high-quality, large-scale classification and regression QSAR models, utilizing publicly available data on SIRT2 inhibitor potency and selectivity. By providing an intuitive, web-based interface, SIRT2i_Predictor is made available to the broader scientific community for use.

Materials and Methods

Dataset preparation

The initial dataset was compiled by extracting records from the ChEMBL database (release 30), focusing on compounds with reported inhibitory activity against SIRT2, with activity measurements either as IC50 values or inhibition percentages (Inh%) [28]. Additional compounds were sourced from patent US20160376238A1 [29], with data extracted using ChemDataExtractor (v 1.3.0) software [30]. The raw dataset was categorized into four groups (Datasets 1–4) based on their intended use (details provided below). Following data collection, all datasets underwent manual curation to remove records that did not pertain specifically to SIRT2 inhibition, excluding activities related to other processes (such as defatty-acylation). The curated datasets were then pre-processed by normalizing SMILES representations, removing duplicate entries, stripping salts, and standardizing molecule structures. RDKit (v 2021.03.4) [31] was used for these tasks. Duplicates were reviewed, and the most representative record for each compound group was kept. When both IC50 and Inh% values were available, the IC50 value was preferred.

For regression models, IC50 values were converted into pIC50 values ($pIC50 = -\log10(IC50)$), while for classification models, compounds were categorized based on IC50 and Inh% thresholds. Compounds were classified as "SIRT2 active" if IC50 ≤ 50 μM, or Inh% ≥ 80% at 200 μM, ≥ 70% at 100 μM, ≥ 60% between 50–100 μM, or ≥ 50% at concentrations below 50 μM. Compounds were assigned to the "SIRT2 inactive" category if IC50 ≥ 90 μM or Inh% ≤ 40% (above 100 μM). For multiclass models, similar classification criteria were applied using data for SIRT1 and SIRT3. To prepare for training, Datasets 1–4 were split into training (70%) and testing (30%) sets using stratified splitting from the scikit-learn library (v 1.1.1) [32]. To address class imbalance, the SMOTE technique from the imbalanced-learn library (v 0.9.1) [33] was used before training the classification models.

Calculation of molecular features and feature selection

Once the datasets were prepared, molecular features were generated by encoding compounds using various fingerprints: 166-bit MACCS keys, 1024-bit extended-connectivity fingerprints (ECFP4 and ECFP6), and 1613 two-dimensional descriptors calculated with the Mordred tool (v 1.2.0) [34]. Descriptors were reduced by removing those with zero or NaN values, followed by standardization. Descriptors with low variance (below 0.1) were removed. Pearson's correlation coefficient was used to detect highly correlated descriptors (above 0.9), and one of each pair was retained. The final set of descriptors for model building was selected through recursive feature elimination with cross-validation (CV) using scikit-learn (v 1.1.1) [32]. A decision tree classifier with 10-fold cross-validation was employed for feature selection. This process was carried out on the training set alone.

Model development and evaluation

This study used five different machine learning (ML) algorithms to develop predictive models: random forest (RF), support vector machines (SVM, including both support vector classification (SVC) and support vector regression (SVR)), k-nearest neighbors (KNN), extreme gradient boosting (XGBoost), and deep neural networks (DNN). The models for RF, SVC, SVR, KNN, and XGBoost were built using scikit-learn and the XGBoost library (v 1.5.1), while DNN models were created using TensorFlow (v 2.9.1) [35]. Hyperparameter optimization for RF, SVC, SVR, KNN, and XGBoost models was carried out using Bayesian optimization with five-fold cross-

validation (CV) from the scikit-optimized library (v 0.8.1). For DNN models, both hyperparameter tuning and network architecture optimization were performed using custom scripts in Keras Tuner (v 1.1.1) [36], utilizing Bayesian optimization with five-fold CV.

The three types of models were trained using the respective datasets: regression models (Dataset 1), binary classification models (Dataset 2), and multiclass classification models (Datasets 3 and 4). Internal validation for regression models included metrics such as the coefficient of determination (R^2), cross-validated correlation coefficient (Q^2), and root mean square error (RMSE) for both training (RMSEint) and cross-validation (RMSECV) [37]. Y-scrambling, a technique for assessing model robustness, was applied by generating 100 models with randomly shuffled data, keeping the same hyperparameters. The external predictive power of regression models was evaluated using metrics including R^2ext (external R^2) (Equation (1)), RMSEext (external RMSE) (Equation (2)) [37], Q^2Fn metrics (Q^2F1, Q^2F2, Q^2F3) (Equations (3)–(5)) [38-40], r^2m metrics (r^2m, $\bar{r}^2$m, Δr^2m) (Equations (6) and (7)) [41, 42], and the concordance correlation coefficient (CCC) (Equation (8)) [43].

$$R^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \overline{y})^2} \quad (1)$$

$$RMSE_{ext} = \sqrt{\frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{n_{EXT}}} \quad (2)$$

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_{TR})^2} \quad (3)$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_{EXT})^2} \quad (4)$$

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2\right]/n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \hat{y}_{TR})^2\right]/n_{TR}} \quad (5)$$

$$r^2_m = r^2\left(1 - \sqrt{(r^2 - r^2_0)}\right) \quad (6)$$

$$\Delta r^2_m = \left|r^2_m - r'^2_m\right| \quad (7)$$

$$CCC = \frac{2\sum_{i=1}^{n_{EXT}} (yi - \overline{y})(\hat{y}i - \overline{\hat{y}})}{\sum_{i=1}^{n_{EXT}} (yi - \overline{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}i - \overline{\hat{y}})^2 + n_{EXT}(\hat{y}i - \overline{\hat{y}})^2} \quad (8)$$

In Equations (1) to (8), TR stands for the training set, and EXT refers to the test set, or external set. The symbol yi represents the actual experimental data, while yi indicates the predicted values. The average of the experimental data is given by $\bar{y}i$, and the average of the predicted data is denoted as y. The coefficients $r20$ and $r2$ represent the determination coefficients of the regression function based on the experimental and predicted values for the external set. r02 is calculated when the regression line is forced through the origin, whereas r2 does not impose this condition. The coefficient r2m is determined by using the experimental values on the y-axis, while $r'2m$ uses the same values on the x-axis. The final value for rm2 is the average of rm2 and r'2m.

$$hi = x_i^T (X^T X)^{-1} xi \quad (9)$$

$$h* = \frac{3(m+1)}{p} \quad (10)$$

The applicability domain of the regression-based models was assessed using the leverage approach [44]. Leverage values (hi) were calculated as per Equation (9), where X represents the matrix of key molecular descriptors from the training set, and xi is the descriptor vector for a query molecule. The threshold value, h*, was determined from

Equation (10), where m is the number of features and p the number of training molecules. Feature importance was evaluated via scikit-learn's permutation importance method, using 30 repetitions.

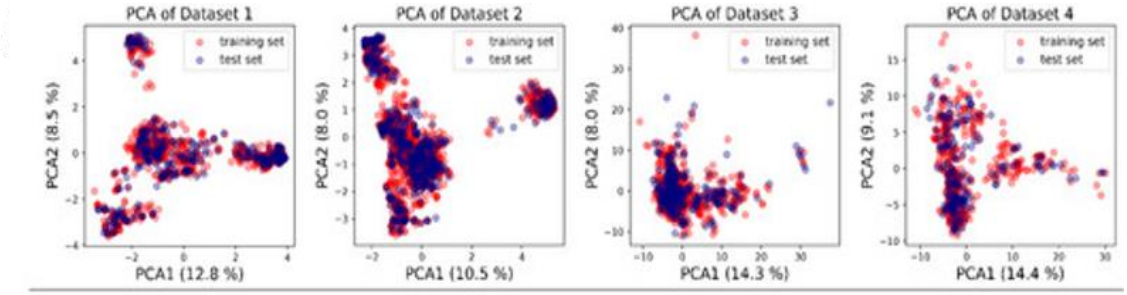
For model evaluation, the following metrics were employed: balanced accuracy, recall, precision, F1-score, Matthews correlation coefficient (MCC), and ROC (receiver operating characteristics) curve area (ROC_AUC). These metrics were derived from confusion matrices that contain values for true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) results. Sensitivity (or recall) was defined as Sensitivity = TP/(TP + FN), and specificity as Specificity = TN/(TN + FP). Balanced accuracy, the average of sensitivity and specificity, is particularly useful for imbalanced datasets and was calculated as BA = (Sensitivity + Specificity)/2. Precision was calculated as Precision = TP/(TP + FP), and the F1-score as the harmonic mean between precision and recall, F1 = 2 × (Precision × Recall)/(Precision + Recall). The MCC provides an overall summary of model performance across all confusion matrix categories. A value greater than 0 for MCC indicates good performance across all categories. The ROC curve plots true-positive rates against true-negative rates at various thresholds, with ROC_AUC quantifying the model's ability to rank randomly chosen positive examples higher than negative ones. For multiclass models, macro-averages were used for ROC_AUC, precision, recall, and F1-score, calculated via a one-vs-rest approach.

External validation of the models was carried out by generating a decoy dataset with the DUD-E server and combining it with an external validation set [45]. For interpreting atomic-level contributions to the model's predictions, similarity maps were created using the RDKit method based on Riniker *et al.*'s approach [46]. Chemical space projections from virtual screening of the SPECS database [47] were visualized using self-organizing maps, as described in earlier work [20].

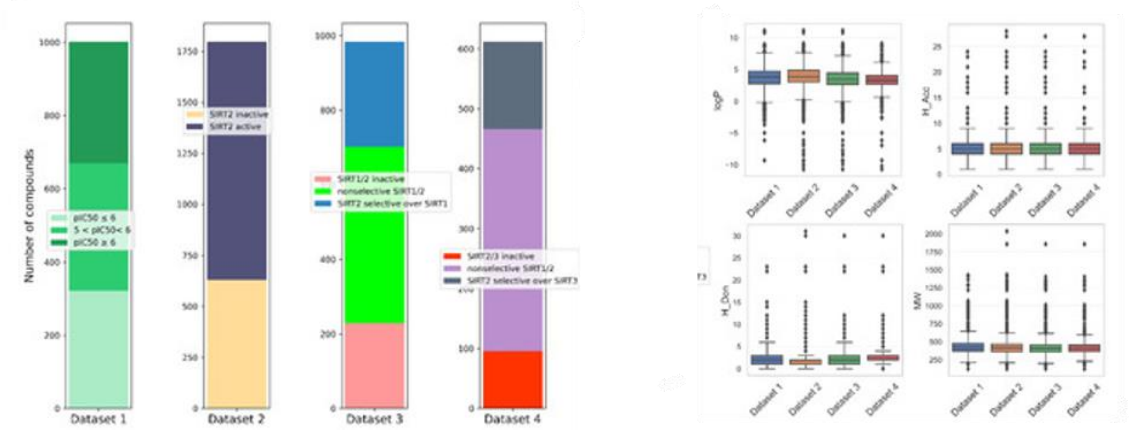
Results and Discussion

Data for model construction

The structural and activity data for SIRT2 inhibitors were gathered from the ChEMBL database and relevant literature (see materials and methods section), yielding 1797 unique records. Due to differences in testing across various isoforms, such as SIRT1 and SIRT3, the initial dataset was divided into four subsets (Datasets 1–4) **(Figures 2 and 3, and Table 1).** Dataset 1 was specifically designed for developing a regression-based QSAR model, while Datasets 2 through 4 were used to construct several types of classification models. The distribution of activity values (pIC50 in Dataset 1) and the classification across different activity categories in Datasets 2–4 are presented in **Figure 2**, with a detailed overview of each dataset's characteristics provided in **Table 1.**



a)



b) c)

Figure 2. Summary Statistics for the Datasets Used in the Analysis. (a) Principal component analysis (PCA) of the chemical space across the datasets. The PCA plots were generated based on the descriptors and fingerprints used in the final machine learning (ML) models. (b) The distribution of data within each dataset. (c) Breakdown of the datasets according to Lipinski's Rule of Five.

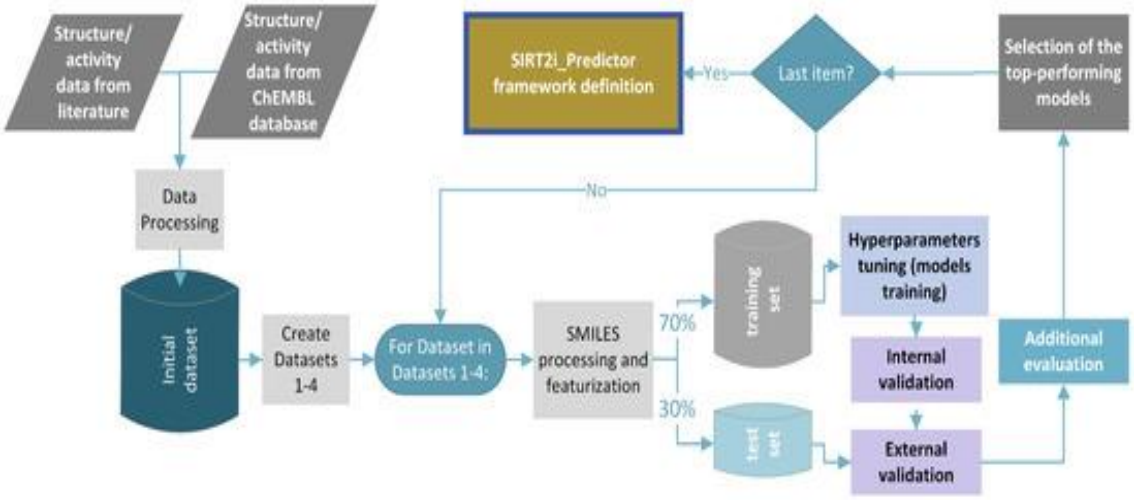


Figure 3. Overview of the Protocol for ML Model Development and Validation.

Table 1. Description of the datasets.

Dataset	Number of Compounds	Measured Activity Metric(s)	Target Sirtuin Protein(s)	Classified Activity Categories
Dataset 1	1002	pIC50	SIRT2	pIC50 (continuous values)
Dataset 2	1797	pIC50 and Inhibition %	SIRT2	Active or Inactive
Dataset 3	984	Inhibition %	SIRT1 and SIRT2	Selective, Non-selective, or Inactive
Dataset 4	612	Inhibition %	SIRT2 and SIRT3	Selective, Non-selective, or Inactive

Dataset 1 was specifically created using compounds with known SIRT2 inhibitory activity, measured as pIC50 values. It included 1002 compounds, with pIC50 values ranging between 4 and 7.96. The largest dataset, Dataset 2, consisted of 1797 compounds, containing both pIC50 and Inh% activity data. Based on the criteria in materials and methods section, the compounds in Dataset 2 were divided into two categories—SIRT2 active and SIRT2 inactive—with approximately one-third of the compounds labeled as inactive **(Figure 2).**

Dataset 3 was composed of compounds that had reported inhibitory activity against both SIRT1 and SIRT2, expressed as either pIC50 or Inh%, while Dataset 4 included compounds with inhibitory data for both SIRT2 and SIRT3, measured as pIC50 or Inh%. Therefore, Datasets 3 and 4 were categorized into three groups: compounds that were inactive against both SIRT1(3) and SIRT2, SIRT2-selective compounds, and non-selective compounds

that inhibited both SIRT1(3) and SIRT2 **(Figure 2 and Table 1).** It is important to note that the bioactivity values in this study were collected from various experimental methods (such as fluorimetric, luminescence, electrophoretic mobility shift, and scintillation counting assays) and experimental conditions (e.g., variations in incubation times and acetyl–lysine substrates with different Km values). To minimize variability due to these factors and clarify the class distinctions, compounds with activity values near the ambiguous boundary between classes were excluded from Datasets 2–4. These compounds, labeled as "twilight zone" compounds, had activity values in the IC50 range of 50–90 μM (for Inh% criteria, refer to materials and methods section). The increasing use of large-scale QSAR models based on diverse ChEMBL datasets has led to studies employing similar data gathering and processing approaches as this study [48-53].

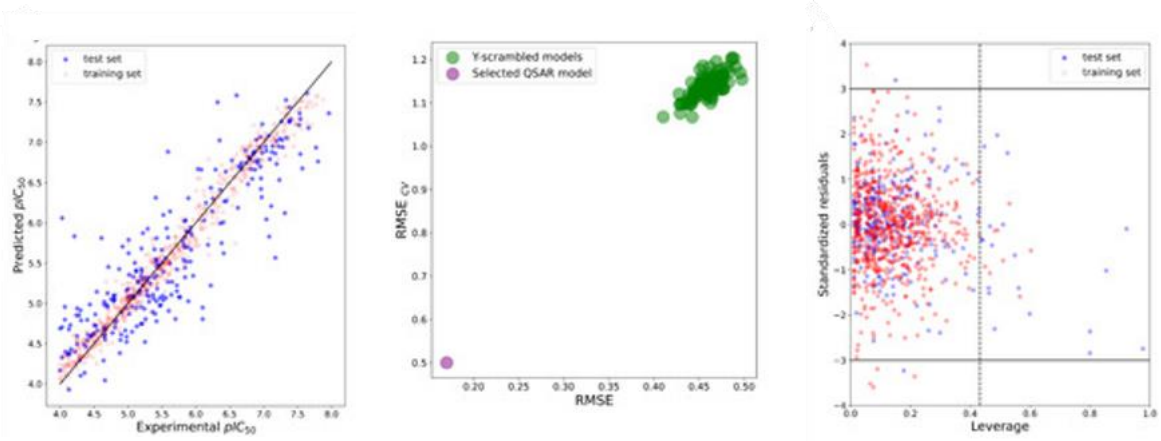
The descriptive analysis of the datasets showed that most of the compounds followed Lipinski's Rule of Five, though some outliers were observed in each dataset **(Figure 2)** [54]. The compounds were pre-processed, and molecular descriptors and fingerprints were generated as described in materials and methods section. Prior to model creation, the datasets were divided into training and test subsets using stratified random sampling (70% for training and 30% for testing), ensuring balanced activity distributions. Principal component analysis (PCA) confirmed that this sampling approach maintained comparable coverage of chemical space between the training and test datasets **(Figure 2)**. To address the slight imbalance in the classification datasets (Datasets 2–4), the SMOTE algorithm was applied to generate synthetic samples of the minority classes before training the classification models.

Model building and evaluation

This work involved creating a range of machine learning models for regression, binary classification, and multiclass classification tasks. These models combined five algorithms—random forest (RF), support-vector machines (SVM, encompassing support-vector classification (SVC) and support-vector regression (SVR)), k-nearest neighbors (KNN), extreme gradient boosting (XGBoost), and deep neural networks (DNN)—with four different molecular representation types (Mordred descriptors, ECFP4 fingerprints, ECFP6 fingerprints, and MACCS keys) (details in materials and methods section). ECFP and MACCS fingerprints stand out in recent cheminformatics literature as the most widely used due to their computational efficiency and proven effectiveness across numerous studies [21]. This popularity and performance record were the key factors influencing their inclusion here. The study's overall pipeline is shown in **Figure 3**. Models specific to each dataset (regression or classification) underwent training with hyperparameter optimization via Bayesian methods and five-fold cross-validation. For the DNN architectures, a broader search covering structural parameters (such as hidden layer count, neuron numbers, and dropout rates) was implemented using Keras Tuner and Bayesian optimization. Initial assessment of the tuned models relied on internal and cross-validation metrics. Generalization capability was then tested on an independent external test set. The highest-performing models were chosen through a consensus strategy, supplemented by task-specific additional evaluations described in the following subsections. These top models ultimately formed the basis of the SIRT2i_Predictor system **(Figure 3).**

Regression models

Global QSAR regression models for Dataset 1 were built by systematically pairing the five algorithms with each of the four molecular feature sets (MACCS, ECFP4, ECFP6 fingerprints, and Mordred descriptors [34]). A feature selection step (materials and methods section) reduced the Mordred set to 52 descriptors for final use, while all fingerprint bits were retained without pruning. Training incorporated Bayesian hyperparameter tuning with five-fold cross-validation. Preliminary quality checks used standard internal metrics: training-set coefficient of determination (R^2), cross-validated Q^2, training RMSE (RMSEint), and cross-validation RMSE (RMSECV). Model acceptance followed the established thresholds of Golbraikh and Tropsha ($R^2 > 0.6$ and $Q^2 > 0.5$) [55]. To rule out overfitting or chance correlations, Y-scrambling was conducted by training 100 models on activity-permuted data. Results confirmed that performance was genuine and the models were trustworthy **(Figure 4).**



a) b) c)

Figure 4. Top Regression-Based Model: XGBoost: ECFP4. (a) Comparison between predicted and experimental pIC50 values. (b) Results from the Y-scrambling test. (c) The model's applicability domain, with the dashed line indicating the leverage threshold (h*).

External validation was carried out to evaluate how the models performed when tested on data outside the training set. In line with the guidelines set by the Organization for Economic Co-operation and Development (OECD), external validation of QSAR models was done by calculating the goodness-of-fit using two main parameters: the coefficient of determination (R^2ext, which should be greater than 0.6) and the root mean square error (RMSEext, where a lower value indicates better performance) [38, 56]. All the models demonstrated similar performance on these metrics, with models using ECFP4 and ECFP6 fingerprints showing a slight edge **(Table 2)**. However, as previous studies have suggested, relying solely on R^2ext can sometimes lead to an overly optimistic view of the model's ability to predict external data. This is because the R^2ext value is influenced by factors such as the response value range and distribution in the test set compared to the training set [43, 57, 58]. As a result, additional evaluation steps were employed to assess the external predictive ability of the QSAR models. These evaluations focused on factors like precision (the variation of observations from the fitting line), accuracy (how close the regression line is to the ideal slope of 1 in the Yobserved vs. Ypredicted plot), and ensuring there was no bias in the response scale [43, 58]. Additional metrics like $\underline{r}^2m$ (where $\underline{r}^2m > 0.5$ and $\Delta r^2m < 0.2$) [41, 42], the Q^2Fn metric, and Concordance Correlation Coefficient (CCC) values with thresholds defined by Chirico and Gramatica (Q^2F1, Q^2F2, $Q^2F3 > 0.7$, CCC > 0.85) were also applied **(Table 2)** [43, 59]. Furthermore, the criteria from Golbraikh and Tropsha were also assessed, including the following: $((R^2-R^2_0)/R^2 < 0.1$ or $(R^2 - R'^2_0)/R^2 < 0.1$, $0.85 \leq k$ (or k') ≤ 1.15, and $|R^2-R'^2_0| < 0.3$) [55]. The majority of the models met almost all of these criteria, though some models failed to meet the Δr^2m and CCC benchmarks **(Table 2)**. After eliminating models that did not meet these criteria, two of the top-performing models were selected: the XGBoost: ECFP4 model **(Figure 4)** and the KNN: ECFP6 model **(Table 2)**.

Table 2. External Validation Parameters for Regression-Based QSAR Models.

ML Algorithm	Molecular Feature	R^2ext	RMSEext	R^2m	ΔR^2m	Q^2F1	Q^2F2	Q^2F3	CCC
RF	Descriptors	0.7	0.55	0.52	0.27	0.7	0.7	0.7	0.81
	ECFP4	0.75	0.5	0.6	0.23	0.75	0.75	0.75	0.85
	MACCS	0.71	0.53	0.55	0.26	0.71	0.71	0.71	0.82
	ECFP6	0.77	0.48	0.62	0.21	0.77	0.77	0.76	0.86
SVR	Descriptors	0.62	0.61	0.44	0.31	0.62	0.62	0.62	0.77
	ECFP4	0.74	0.51	0.63	0.13*	0.74	0.74	0.73	0.84
	MACCS	0.68	0.57	0.55	0.21	0.68	0.68	0.68	0.81
	ECFP6	0.74	0.51	0.63	0.18*	0.74	0.74	0.74	0.86
XGBoost	Descriptors	0.67	0.58	0.53	0.25	0.68	0.68	0.68	0.82
	ECFP4	0.75	0.5	0.64	0.17	0.74	0.74	0.74	0.86

		(0.79)b	(0.46)b	(0.7)b	(0.17)*,b	(0.75)b	(0.75)b	(0.74)b	(0.86)a,b
	MACCS	0.71	0.53	0.58	0.24	0.7	0.7	0.7	0.82
	ECFP6	0.73	0.52	0.62	0.2	0.73	0.73	0.73	0.87
KNN	Descriptors	0.68	0.56	0.56	0.23	0.68	0.68	0.68	0.86
	ECFP4	0.74	0.51	0.64	0.13*	0.74	0.74	0.74	0.87
	MACCS	0.6	0.63	0.47	0.16*	0.6	0.6	0.6	0.79
	ECFP6	0.76 (0.77)b	0.49 (0.48)b	0.66 (0.68)b	0.12 (0.11)*,b	0.76 (0.76)b	0.76 (0.76)b	0.76 (0.76)b	0.87 (0.87)a,b
DNN	Descriptors	0.66	0.58	0.57	0.03*	0.66	0.66	0.66	0.81
	ECFP4	0.74	0.51	0.63	0.18*	0.73	0.73	0.73	0.84
	MACCS	0.68	0.56	0.56	0.16*	0.68	0.68	0.67	0.80
	ECFP6	0.73	0.52	0.63	0.17*	0.73	0.73	0.73	0.81
Criteria		>0.6		>0.5	<0.2	>0.7	>0.7	>0.7	>0.85

Validation of regression models

In accordance with OECD guidelines, it is crucial to establish the boundaries of chemical space within which the model can make reliable predictions. This is referred to as the applicability domain (AD) of the QSAR model, which should be clearly defined during external validation. These boundaries must be considered when predicting unknown compounds [38, 44, 56]. One widely applied method to estimate these boundaries in regression QSAR models is the leverage method [44]. Leverage values represent how far each compound is from the central point (centroid) of the training set in the feature space. These values are typically displayed in Williams plots, which plot leverage against standardized residuals, making it easy to identify compounds that exceed the thresholds for leverage or residuals **(Figure 4).**

For two of the top-performing models (XGBoost: ECFP4 and KNN: ECFP6), a number of compounds from the test set were found to be outside the applicability domain **(Table 2 and Figure 4)**. The KNN: ECFP6 model had a more constrained coverage of chemical space, with 39 test compounds outside the AD, while the XGBoost:ECFP4 model only had 24 compounds falling outside its boundaries. When compounds outside the AD were excluded, the performance of the XGBoost: ECFP4 model improved substantially, whereas the KNN:ECFP6 model showed minimal improvement **(Table 2)**. This suggests that the KNN:ECFP6 model has less predictive power and narrower chemical space coverage within the AD. Additionally, the KNN: ECFP6 model exhibited lower robustness during internal validation.

Given that global QSAR models are intended to cover a wider range of chemical space and provide robust and accurate external predictions, the XGBoost: ECFP4 model was selected for further development. However, it is important to note that the regression models were trained on Dataset 1, which consists mostly of active compounds (914 active compounds, compared to only 88 compounds in the "twilight zone" (IC50 = 50–90 µM) or "inactive" (IC50 > 90 µM) categories) (materials and methods section). Therefore, these models are primarily suitable for predicting the pIC50 values of active compounds or compounds identified as likely active by classification models. This limitation is discussed in more detail in Section "SIRT2i_predictor framework for discovering novel inhibitors". Furthermore, the diversity in data sources contributing to the ChEMBL database may lead to inconsistencies that affect the prediction accuracy of the regression models [48]. For this reason, classification-based models, which do not rely on these issues, might offer better performance in identifying active compounds.

Binary classification models

As outlined in the study protocol **(Figure 3)**, five different machine learning algorithms, in combination with four molecular features, were tested using Bayes hyperparameter optimization and five-fold cross-validation (CV) to develop binary classification models with Dataset 2. The goal was to train models to classify compounds as SIRT2 inhibitors or inactive compounds. The criteria for assigning compounds to these categories are provided in materials and methods section. Before training, the Mordred descriptors underwent a feature-selection process (see materials and methods section). The final modeling used 233 selected Mordred descriptors, and molecular fingerprints were utilized without further reduction in the number of bits.

The models' internal predictive power, stability, and robustness were assessed using internal validation metrics. The performance of the binary classification models was evaluated using an external test set, where the following

metrics were used: balanced accuracy (BA), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (ROC_AUC), precision, recall, and F1-score **(Table 3)**. In general, the algorithms performed similarly on the external test set, with the RF, SVC, and DNN models, especially those using descriptors, showing slightly better external predictive power.

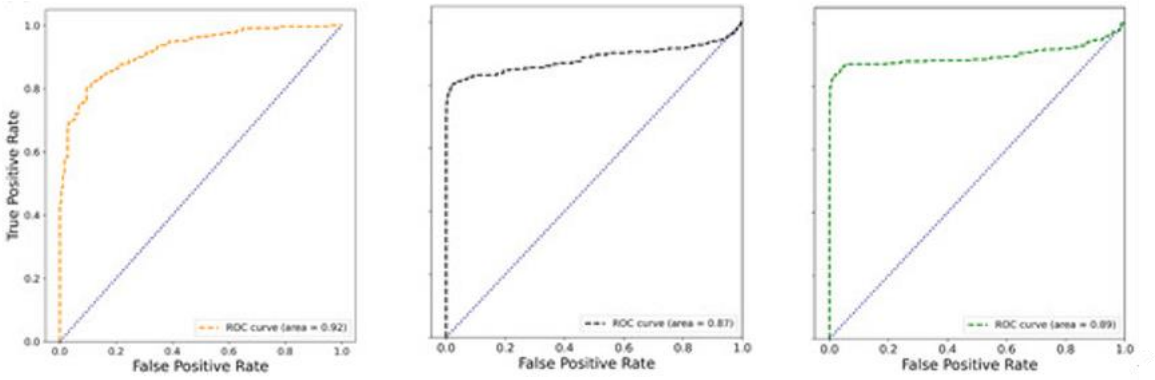
Table 3. Parameters for external validation of the binary classification models.

ML Algorithm	Molecular Feature	BA	MCC	ROC_AUC	Precision	Recall	F1
RF	Descriptors	0.88	0.74	0.94	0.86	0.88	0.87
	ECFP4	0.84	0.66	0.92	0.82	0.84	0.83
	MACCS	0.82	0.62	0.91	0.8	0.82	0.81
	ECFP6	0.85	0.68	0.92	0.83	0.85	0.84
SVR	Descriptors	0.88	0.74	0.95	0.87	0.88	0.87
	ECFP4	0.81	0.63	0.9	0.82	0.81	0.82
	MACCS	0.8	0.59	0.87	0.79	0.8	0.79
	ECFP6	0.79	0.62	0.9	0.83	0.79	0.81
XGBoost	Descriptors	0.86	0.72	0.94	0.85	0.86	0.85
	ECFP4	0.81	0.62	0.91	0.8	0.81	0.81
	MACCS	0.8	0.60	0.9	0.8	0.8	0.8
	ECFP6	0.81	0.62	0.91	0.81	0.81	0.81
KNN	Descriptors	0.79	0.56	0.88	0.77	0.79	0.77
	ECFP4	0.82	0.62	0.9	0.8	0.82	0.81
	MACCS	0.82	0.62	0.88	0.8	0.82	0.81
	ECFP6	0.84	0.65	0.91	0.81	0.84	0.82
DNN	Descriptors	0.89	0.75	0.94	0.85	0.86	0.86
	ECFP4	0.83	0.65	0.91	0.8	0.81	0.8
	MACCS	0.8	0.58	0.89	0.8	0.8	0.8
	ECFP6	0.82	0.64	0.9	0.79	0.82	0.8

[a] This paper reports the average performance values for each class.

In the field of virtual screening (VS), the aim is often to identify active molecules from vast databases that are predominantly filled with inactive compounds. Machine learning (ML) models used in VS are considered more effective if they can cover a wider chemical space, as the primary goal is to uncover novel and diverse active compounds. A binary model trained on the largest dataset, Dataset 2, is expected to offer the broadest chemical space coverage, thus making it more advantageous for VS tasks compared to those trained on smaller datasets. To further test the relevance of these models for practical applications, a real-life VS simulation was carried out by generating almost 20,000 virtual decoy molecules, which were assigned to the inactive class. These decoys were crafted to be structurally dissimilar from known active compounds while keeping similar physical properties. The decoy dataset was then combined with an external set to create an imbalanced database with a 1:40 ratio of active to inactive compounds. The models were evaluated on their ability to retrieve the active molecules.

In these conditions, statistical parameters were re-calculated by incorporating early enrichment metrics **(Table 4).** These metrics are crucial because early identification of active compounds is a key aspect of VS, where only the top-ranked compounds are selected for experimental testing. Early recognition is an important indicator of a model's ability to identify active compounds quickly in the ranking process. We applied ROC EF 0.5%, 1%, 2%, and 5%, which assess the coverage area under the curve for 0.5%, 1%, 2%, and 5% of false positives, respectively [20, 60]. Given that the decoy set contained a far greater number of chemically diverse inactive compounds, the RF: ECFP4 binary model stood out as having the strongest predictive performance **(Table 4 and Figure 5)**. In the highly imbalanced decoy dataset, the RF: ECFP4 model not only demonstrated better sensitivity, specificity, precision, and overall robustness, but it also excelled in early recognition. With only 0.5% false positives, this model managed to retrieve more than 70% of the true active molecules **(Table 4).** It is noteworthy that most of the inactive compounds in Dataset 2 were chemically similar to the active ones, whereas the decoy dataset was enriched with compounds that were topologically distinct. Since the decoy dataset provided a more reliable assessment of model performance in VS settings, the RF: ECFP4 model was chosen for further analysis.



a) b)

Figure 5. ROC curves for the leading RF: ECFP4 binary model. (a) Results of the external validation; (b) Validation results using the decoy dataset; (c) Validation results after applicability-domain modifications on the decoy dataset.

Table 4. Performance metrics for the binary models tested on the decoy dataset.

ML Algorithm	Molecular Feature	BA	MCC	ROC_ AUC	Precision (a)	Recall (a)	F1 (a)	EF05%	EF1%	EF2%	EF5%
Random Forest (RF)	Descriptors	0.68	0.09	0.87	0.51	0.68	0.35	0.63	0.67	0.68	0.73
	ECFP4	0.81 (0.9) b	0.19 (0.52) b	0.87 (0.89) b	0.53 (0.67) b	0.81 (0.9) b	0.49 (0.73) b	0.74 (0.74) b	0.74 (0.74) b	0.76 (0.76) b	0.77 (0.8) b
	MACCS	0.66	0.08	0.82	0.51	0.66	0.35	0.55	0.56	0.59	0.62
	ECFP6	0.75	0.14	0.87	0.52	0.75	0.43	0.72	0.74	0.76	0.78
Support Vector Regression (SVR)	Descriptors	0.69	0.1	0.89	0.51	0.69	0.36	0.43	0.56	0.62	0.71
	ECFP4	0.46	−0.06	0.8	0.48	0.46	0.05	0.75	0.75	0.75	0.76
	MACCS	0.62	0.06	0.83	0.51	0.62	0.32	0.39	0.61	0.68	0.74
	ECFP6	0.47	−0.07	0.8	0.47	0.47	0.03	0.76	0.76	0.77	0.77
XGBoost	Descriptors	0.71	0.11	0.85	0.51	0.71	0.39	0.41	0.44	0.48	0.54
	ECFP4	0.74	0.13	0.87	0.52	0.74	0.42	0.35	0.39	0.43	0.52
	MACCS	0.64	0.07	0.73	0.51	0.64	0.32	0	0	0.02	0.2
	ECFP6	0.71	0.11	0.85	0.51	0.71	0.39	0.37	0.38	0.44	0.5
K-Nearest Neighbors (KNN)	Descriptors	0.66	0.08	0.76	0.51	0.66	0.37	0.09	0.23	0.26	0.29
	ECFP4	0.72	0.12	0.8	0.52	0.72	0.41	0	0	0	0
	MACCS	0.64	0.07	0.75	0.51	0.64	0.33	0	0	0	0
	ECFP6	0.72	0.11	0.8	0.52	0.72	0.41	0	0	0	0
Deep Neural Networks (DNN)	Descriptors	0.72	0.12	0.8	0.51	0.71	0.38	0	0	0	0
	ECFP4	0.73	0.13	0.84	0.52	0.73	0.43	0.1	0.25	0.32	0.41
	MACCS	0.69	0.1	0.79	0.51	0.62	0.29	0.04	0.08	0.17	0.23
	ECFP6	0.67	0.09	0.81	0.51	0.67	0.38	0.17	0.25	0.34	0.43

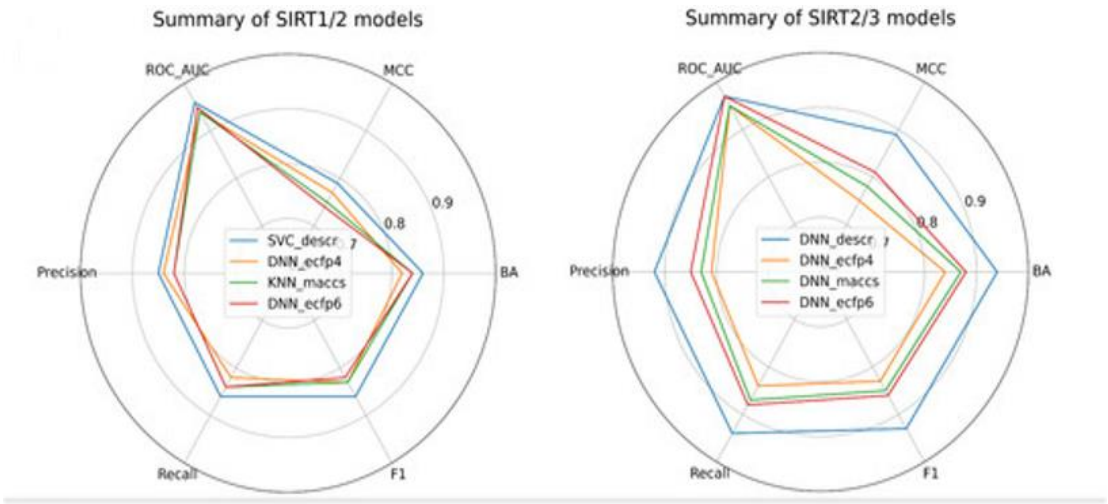
[a] The table shows average values for all classes. b The values in parentheses indicate the performance after applicability domain adjustments.

The applicability domain for the selected models was established using the indeterminate-zone approach [61-63]. Predictions that fall within this "indeterminate zone" (in-zone predictions) are deemed uncertain, whereas those outside this zone are considered more reliable. For binary models, the indeterminate zone was set as a prediction probability range of 0.5 ± 0.1 for each class. When applicability-domain corrections were applied, the RF: ECFP4 model's performance showed substantial improvement **(Table 4 and Figure 5).**

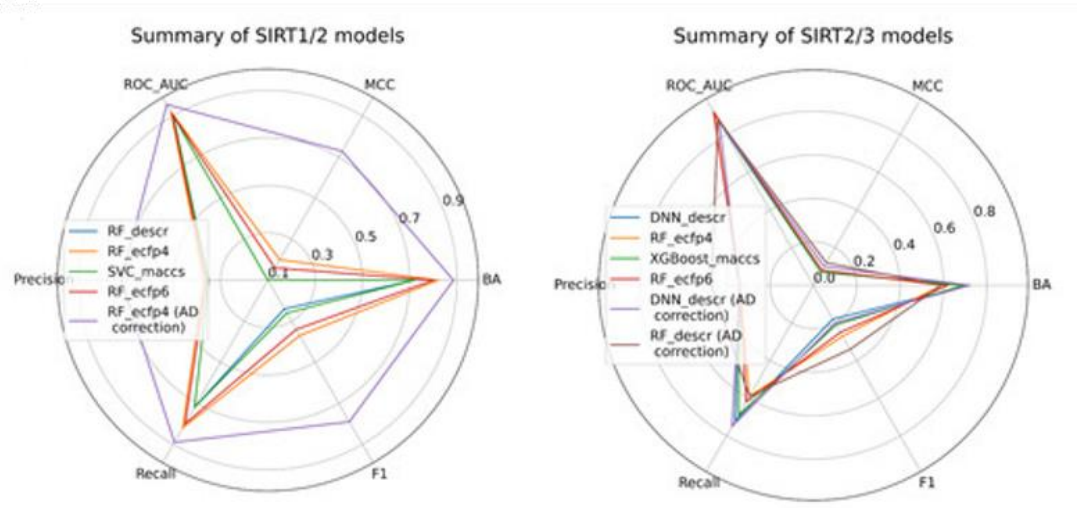
Multiclass classification models

SIRT1 and SIRT3 are the closest homologues of SIRT2, all grouped in Class I, which complicates the development of selective inhibitors for SIRT2 [64]. Recent studies have linked the safety of SIRT2 inhibitors to their selectivity, making selectivity a key factor in the design of novel SIRT2 inhibitors [19]. A large portion of the structure-activity data in ChEMBL involves molecules with activity against multiple sirtuin isoforms, with a particular focus on SIRT1 and SIRT3. However, the subset of SIRT2 inhibitors with activity data for SIRT1 or SIRT3 is much smaller and more imbalanced, which could hinder the effectiveness of models trained on these data **(Figure 2).** The primary objective of this section was to create and validate models that predict the selectivity of potential inhibitors.

As per the established protocol **(Figure 3),** a variety of machine learning (ML) algorithms and molecular features were used to construct and assess selectivity models. Two separate models were built: one for predicting SIRT1/2 selectivity and the other for SIRT2/3 selectivity. Due to the limited number of compounds with data for all three isoforms and the significant class imbalances, a combined SIRT1/2/3 model was not developed. For the SIRT1/2 selectivity model, 270 Mordred descriptors were selected, while 316 descriptors were chosen for the SIRT2/3 model. No reduction in the number of bits was applied to the fingerprints. These models were designed to classify compounds into different groups: selective SIRT2 inhibitors, non-selective SIRT1/2 or SIRT2/3 inhibitors, and inactive compounds **(Figure 2)**. As with the binary VS models, both external and internal validations were conducted using the same statistical parameters **(Figure 6)**. Internal validation results showed strong predictive performance of the trained models. Interestingly, models using molecular descriptors outperformed those based on fingerprints in external validation with the test set, particularly for Random Forest (RF), Deep Neural Networks (DNN), and Support Vector Classifier (SVC) approaches. This suggests that the physicochemical properties encoded in the descriptors were more influential in determining selectivity than the structural features provided by the fingerprints. Notably, the DNN models showed the best overall performance, especially for the SIRT2/3 selectivity model, indicating that deep learning models were better able to leverage the limited training data available in Dataset 4, the smallest dataset **(Figure 6a).**



a)



b)

Figure 6 illustrates the predictive performance summary for multi-class selectivity models, highlighting the best-performing model for each feature type. Panel (a) shows parameters from the external (test) set validation, while Panel (b) displays parameters from the decoy set validation. Precision, recall, and F1 scores across all panels are presented as macro averages.

The practical utility of these selectivity models could be realized by using them to predict outcomes for a large number of inactive compounds. However, the limited chemical space coverage of the true inactive compounds within Datasets 3 and 4 might restrict how broadly these models can be applied. To better simulate real-world usage against many inactive compounds, the models were additionally tested on decoy datasets (with an active-to-inactive ratio of 1:40), similar to how binary models are evaluated. This enrichment of the inactive class in the decoy set caused a minor shift in the models' predictive metrics.

Interestingly, the analysis using the decoy set showed that ECFP4 molecular representations offered an advantage for the SIRT1/2 models **(Figure 6b)**. Following the decoy set evaluation, the RF: ECFP4 SIRT1/2 model was clearly superior to the other tested models. A similar pattern emerged for the SIRT2/3 models, where the RF: ECFP4 SIRT2/3 model also showed improved predictive accuracy on the decoy set. However, for the SIRT2/3 models, the results were less definitive, as the DNN:descriptors SIRT2/3 model performed comparably **(Figure 6b)**. It is important to note that the SIRT2/3 models generally performed poorly on the decoy sets that were specifically constructed to be highly imbalanced by maximizing the 2D topological dissimilarity among the decoy compounds. This suggests the utility of the SIRT2/3 models may be limited only to compounds that are topologically similar to those already known to be active. The generally weaker performance of the SIRT2/3 models on the decoy dataset is likely due to the restricted size and chemical diversity of the training data used.

To further investigate the viability of these chosen models, applicability domains (AD) were established. Mirroring the approach used for binary models, the indeterminate-zone method was employed to define the AD for the selectivity models. Given that these models yield three class outcomes, the AD for confident predictions was defined as having a probability greater than 0.5 for the predicted class. When only data points falling within this AD were considered, the SIRT1/2 model experienced the most significant boost in predictive statistics **(Figure 6b)**. In contrast, the two promising SIRT2/3 models that performed similarly on the decoy set (RF: ECFP4 SIRT2/3 and DNN:descriptors SIRT2/3) saw only minor statistical improvements after AD refinement **(Figure 6b)**. However, the DNN model achieved substantially better coverage, encompassing nearly 19,000 compounds within the AD boundaries, compared to about 9,000 for the RF model, which led to its selection for subsequent work.

In conclusion, the SIRT1/2 model demonstrated superior predictive capability, while the SIRT2/3 models exhibited lower quality when assessing compounds with topological differences. The primary reason for this difference is probably related to the variation in the size of the datasets used. Considering these limitations regarding dataset scale and diversity, alongside model quality, the SIRT1/2 and SIRT2/3 selectivity models are best suited to serve as secondary tools for analyzing the selectivity of virtual screening results that have already been predicted by the more accurate binary models. Any conflicting predictions—for instance, when a binary

model classifies a compound as active, but a selectivity model deems it inactive—that occur within the models' applicability domains should be approached cautiously. In such conflicting scenarios, the selectivity models (particularly the SIRT1/2 model) can be utilized as an additional means to confirm a compound's activity.

SIRT2i_predictor framework for discovering novel inhibitors

To enhance accessibility and promote the optimal use of the developed models, a comprehensive framework for predicting the activity and selectivity of novel compounds was designed. This framework was integrated into a Python-based application called SIRT2i_Predictor. The structure of SIRT2i_Predictor, as shown in **Figure 7**, comprises four main components: (1) a module selector, (2) a SMILES preprocessor, (3) predictors, and (4) analyzers. In addition to this, a user-friendly web-based graphical user interface (GUI) was developed to make SIRT2i_Predictor more accessible to the broader scientific community **(Figure 8).**

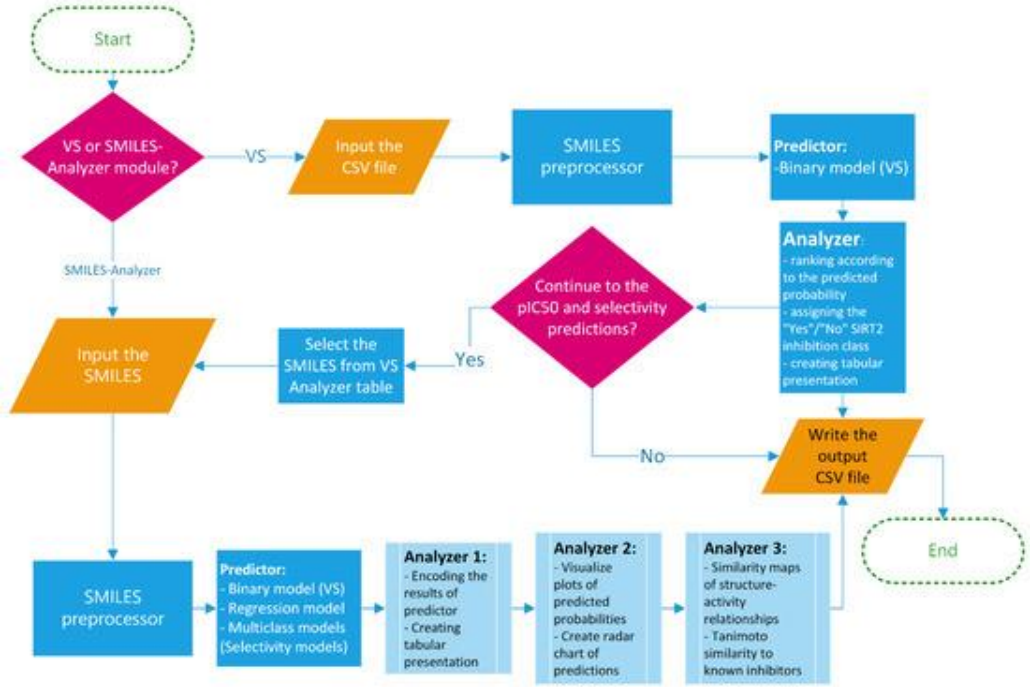


Figure 7. SIRT2i_Predictor framework.

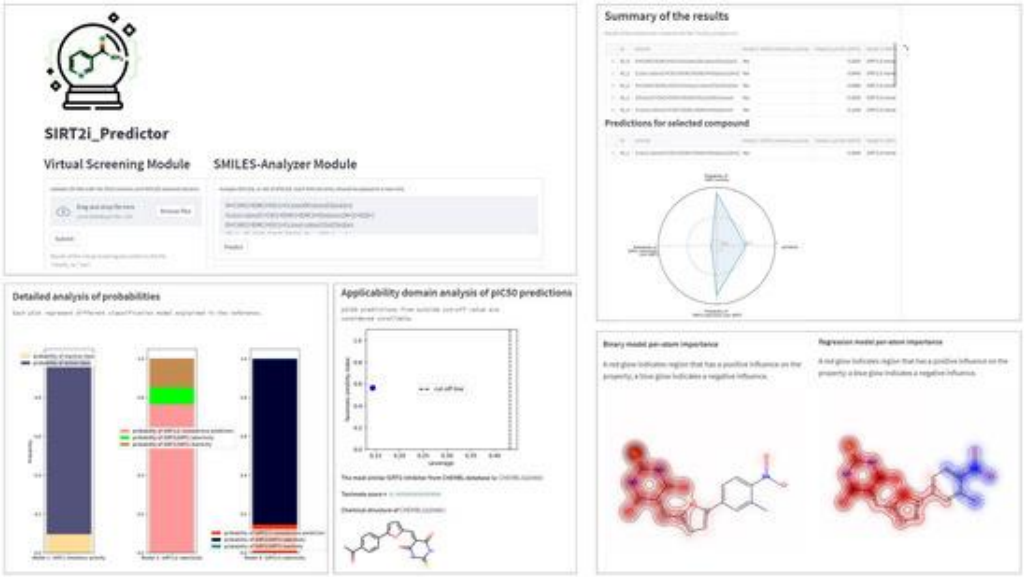


Figure 8. Overview of the core features of SIRT2i_Predictor**.**

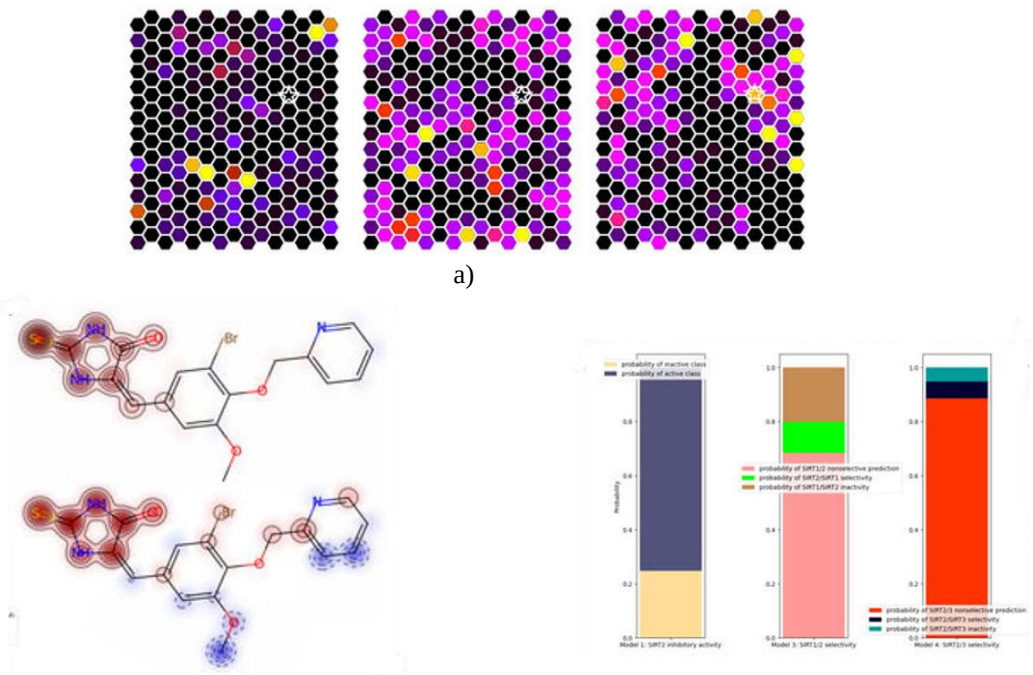
The VS module accepts input via CSV files containing compounds, optionally with associated compound IDs, with a size limit of 200 MB. These SMILES strings are automatically prepared for prediction by a dedicated SMILES pre-processor. The predictor within the VS module is based exclusively on the binary-classification RF: ECFP4 model**.** The selectivity models, particularly the SIRT2/3 model, showed restricted usefulness on the decoy set compared to the binary model, which is likely due to the restricted size and lack of diversity in their respective training datasets (refer to Section "Multiclass classification models"). Conversely, regression models, which were predominantly trained on active compounds, might not be the optimal choice for general VS tasks. Given the scarcity of inactive compounds in the training set (as detailed in Section "Validation of regression models"), the regression model could, however, serve as a useful analytical instrument for a focused examination of compounds previously flagged as active by the binary model. Consequently, the binary RF: ECFP4 classification model was chosen as the primary virtual-screening model because it utilized the largest and most varied dataset and demonstrated superior performance in practical application. The output from the VS module is a generated CSV file detailing the screened molecules, along with their predicted probabilities and class assignments ("Yes" if predicted as an inhibitor, "No" otherwise) **(Figure 8)**. This capability can significantly aid in prioritizing compounds and reducing expenses in extensive *in vitro* screening efforts.

The second component, the SMILES-Analyzer, is designed for a deeper examination of the results generated by the VS module, although it can be utilized independently to analyze specific compounds of interest. This module requires a list of SMILES strings to be manually entered by the user into an input field **(Figure 8)**. The predictor here incorporates all four established Machine Learning models: the binary RF: ECFP4 model, the regression XGBoost: ECFP4 model, the selectivity RF:ECFP4 SIRT1/2 model, and the DNN:descriptors SIRT2/3 model. Different analyzers then provide the user with comprehensive reports on the predicted potencies and selectivity levels of the compounds identified as potential inhibitors **(Figure 7)**. The fundamental purpose of the SMILES-Analyzer is to facilitate the analysis and final selection of VS results slated for subsequent experimental validation. Similar to the VS module, the SMILES-Analyzer can process a large volume of compounds. However, the necessity for each model to receive specific input formats means that the preprocessing step in this module demands greater computational time, which could pose a challenge when screening very extensive databases. Beyond generating textual and numerical summaries of the predictions from all four ML models, the SMILES-Analyzer also facilitates an in-depth, individual inspection of each compound through graphical representations **(Figure 8)**. These graphical interpretations include: a radar chart summarizing the predictions from all four models; a histogram of predicted probabilities to gauge the confidence level (applicability domain) of the classification model outputs; a leverage plot to assess the confidence (applicability domain) of the pIC50\text{pIC}_{50}pIC50 predictions; and maps illustrating the per-atom contributions (both positive and negative) to the predicted SIRT2 activity (derived from both regression and binary models) **(Figure 8)**. These atomic-contribution maps are based on the similarity maps concept, where the impact of each atom is quantified by comparing the predicted probabilities before and after removing the bits in the fingerprint corresponding to that atom [46]. Furthermore, the Graphical User Interface (GUI) provides predictions for the most chemically similar compounds within the ChEMBL database, calculated using Tanimoto similarity and ECFP4 fingerprints. This Tanimoto analysis is included specifically to help evaluate the confidence of the SIRT2/3 selectivity predictions, as the DNN:descriptors SIRT2/3 model exhibited restricted applicability on the chemically diverse decoy set (see Section "Multiclass classification models"). Consistent with the VS module, the SMILES-Analyzer module also produces a final tabular report (as a CSV file) detailing the predictions for all molecules entered.

Benchmarking SIRT2i_predictor against structure-based virtual screening (SBVS)

In a recent study, our team introduced an innovative structure-based virtual screening (SBVS) method that utilizes multiple conformational states of SIRT2, discovered through intensive simulations of the binding-pocket dynamics [20]. By employing different binding-pocket conformations, this approach outperformed traditional single-structure methods in terms of validation metrics and expanded the chemical space coverage for virtual hits. To evaluate whether SIRT2i_Predictor could similarly expand the chemical space of virtual hits, we repeated the prospective SBVS campaign from the previous study, which involved screening approximately 200,000 compounds from the SPECS database [47]. Chemical space coverage was analyzed using self-organizing maps and compared between the two methods. The binary models in SIRT2i_Predictor notably expanded the chemical space of virtual hits, surpassing the chemical space covered by known SIRT2 inhibitors **(Figure 9a)**. While the chemical space expansion was slightly less than that achieved by SBVS, SIRT2i_Predictor showed strong

potential as a comparable tool for identifying novel inhibitor scaffolds. To explore regions of chemical space that were not covered by either the SBVS or ChEMBL datasets, we selected one of the highest-ranked compounds that occupied an unexplored area. This compound, a thiohydantoin derivative, was predicted by SIRT2i_Predictor to be a non-selective SIRT2 inhibitor, with all probability values within the applicability domain **(Figure 9b)**. Analyzing the structure-activity relationship using similarity maps revealed that the thiohydantoin scaffold played a key role in the predicted activity **(Figure 9c)**. Interestingly, no thiohydantoin derivatives had been previously identified as sirtuin inhibitors, further confirming the ability of SIRT2i_Predictor to discover new inhibitor scaffolds and generate structure-activity hypotheses. Moreover, SIRT2i_Predictor's VS module demonstrated a significant time advantage, screening 200,000 compounds in just minutes, in contrast to the SBVS method, which took several hours.



a) b) c)

Figure 9. Comparison of SIRT2i_Predictor and the multi-structure SBVS technique. (a) The chemical space coverage of ChEMBL SIRT2 inhibitors (left), SBVS virtual hits (middle), and SIRT2i_Predictor virtual hits (right) are shown side by side; (b) assessment of the probabilities for the virtual hits originating from a unique portion of chemical space identified by SIRT2i_Predictor (marked with a star in (a)); (c) structure-activity relationships derived from similarity maps for both the binary model (upper plot) and regression model (lower plot) for the virtual hits from this unique chemical space (star marked in (a)). Positive regions contributing to activity are highlighted in red, while negative regions are shown in blue.

In the SBVS experiment, nine molecules identified from previously uncharted chemical spaces displayed activity against SIRT2. However, the IC50 values for two of the lead compounds, as well as the Inh% of five other compounds, fell into the "twilight zone" (IC50 = 50–90 μM; Inh%@200 μM = 40–80%) in the binary model, and only two compounds were unequivocally inactive. To evaluate how SIRT2i_Predictor would perform with compounds from this "twilight zone," the same nine molecules were re-analyzed using the SIRT2i_Predictor virtual screening module. The model's predictions were in line with experimental results, with SIRT2i_Predictor forecasting that none of the compounds would have an IC50 below 50 μM—the threshold for active compounds in the binary models. Five of the compounds were marked as being outside the applicability domain, which corresponds to their placement in the "twilight zone" due to their origin from less explored chemical regions. The remaining four compounds were predicted to be inactive.

Further evaluation of SIRT2i_Predictor was carried out using an in-house database [65, 66]. Unfortunately, predictions for all compounds indicated that they either fell outside the applicability domain or were classified as inactive. To further validate the performance of SIRT2i_Predictor, we compared these predictions with those from an earlier SBVS model. Four compounds that were predicted to be active by the SBVS model were subsequently

tested in vitro. Although some inhibitory activity against SIRT2 was observed, none of the compounds proved to be potent inhibitors, with three falling within the "twilight zone" and one being inactive. These results were consistent with SIRT2i_Predictor's predictions, which aligned with the experimental outcomes, classifying these compounds as either inactive or outside the applicability domain. As a positive control, we used EX-527 (IC50 (SIRT2) = 20 μM) [67], which was not included in the training set for SIRT2i_Predictor. The model effectively differentiated EX-527 from the in-house compounds, and the predictions matched the experimental results.

To sum up, SIRT2i_Predictor was shown to be effective at filtering out less potent compounds while offering a comparable chemical space coverage to the more computationally expensive SBVS methods. The benchmarking results suggest that SIRT2i_Predictor can serve as a valuable addition to SBVS tools, functioning both as a standalone virtual screening tool and as a rapid, convenient filtering method for prioritizing compounds after virtual or in vitro screening, assisting in the selection of the most promising candidates for further biological testing.

Conclusion

SIRT2 inhibitors show significant potential in treating age-related diseases, and preclinical data continues to support their development. However, despite increasing interest, no SIRT2 inhibitors have reached clinical trials. The absence of large-scale, reliable structure–activity relationship (SAR) models for predicting SIRT2 inhibitor potency and selectivity remains a major challenge. Such models could dramatically reduce the time and costs associated with developing new inhibitors. To address this, we compiled all available SAR data and built a set of high-quality machine-learning models for predicting the potency and selectivity of SIRT2 inhibitors. After extensive validation, four models were identified as top performers: the binary RF: ECFP4 and regression XGBoost: ECFP4 models for potency prediction, and the RF: ECFP4 SIRT1/2 and DNN:descriptors SIRT2/3 models for selectivity prediction.

To facilitate practical application, we developed the SIRT2i_Predictor, a Python-based tool featuring an intuitive web interface. The tool enables fast processing of SMILES input and can efficiently evaluate large compound databases for SIRT2 inhibitory potency and SIRT1–3 selectivity, which is particularly useful for virtual screening campaigns and prioritizing compounds for expensive in vitro studies. It also offers visualization tools to highlight the molecular features contributing to activity, making SIRT2i_Predictor an asset in lead optimization efforts. Our benchmarking study indicated that SIRT2i_Predictor complements the SBVS method recently published. The code for database curation, model training, and GUI development is adaptable for other pharmacologically relevant targets, contributing to the creation of broader in silico platforms, a direction we aim to pursue in future work.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Finkel T, Deng CX, Mostoslavsky R. Recent progress in the biology and physiology of sirtuins. Nature. 2009;460:587-91.
2. Haigis MC, Sinclair DA. Mammalian sirtuins: Biological insights and disease relevance. Annu Rev Pathol Mech Dis. 2010;5:253-95.
3. Saunders LR, Verdin E. Sirtuins: Critical regulators at the crossroads between cancer and aging. Oncogene. 2007;26:5489-504.
4. Wang Y, Yang J, Hong T, Chen X, Cui L. SIRT2: Controversy and multiple roles in disease and physiology. Ageing Res Rev. 2019;55:100961.

5. Zhang H, Dammer EB, Duong DM, Danelia D, Seyfried NT, Yu DS, et al. Quantitative proteomic analysis of the lysine acetylome reveals diverse SIRT2 substrates. Sci Rep. 2022;12:3822.
6. de Oliveira RM, Sarkander J, Kazantsev A, Outeiro T. SIRT2 as a therapeutic target for age-related disorders. Front Pharmacol. 2012;3:82.
7. Hong JY, Lin H. Sirtuin modulators in cellular and animal models of human diseases. Front Pharmacol. 2021;12:735044.
8. Zhang L, Kim S, Ren X. The clinical significance of SIRT2 in malignancies: A tumor suppressor or an oncogene? Front Oncol. 2020;10:1721.
9. Jing H, Hu J, He B, Negron Abril YL, Stupinski J, Weiser K, et al. A SIRT2-selective inhibitor promotes c-Myc oncoprotein degradation and exhibits broad anticancer activity. Cancer Cell. 2016;29:297-310.
10. Nielsen AL, Rajabi N, Kudo N, Lundø K, Moreno-Yruela C, Bæk M, et al. Mechanism-based inhibitors of SIRT2: Structure–activity relationship, X-ray structures, target engagement, regulation of α-tubulin acetylation and inhibition of breast cancer cell migration. RSC Chem Biol. 2021;2:612-26.
11. Wawruszak A, Luszczki J, Czerwonka A, Okon E, Stepulak A. Assessment of pharmacological interactions between SIRT2 inhibitor AGK2 and paclitaxel in different molecular subtypes of breast cancer cells. Cells. 2022;11:1211.
12. Karwaciak I, Sałkowska A, Karaś K, Sobalska-Kwapis M, Walczak-Drzewiecka A, Pułaski Ł, et al. SIRT2 contributes to the resistance of melanoma cells to the multikinase inhibitor dasatinib. Cancers. 2019;11:673.
13. Cheng WL, Chen KY, Lee KY, Feng PH, Wu SM. Nicotinic-NAChR signaling mediates drug resistance in lung cancer. J Cancer. 2020;11:1125-40.
14. Hamaidi I, Zhang L, Kim N, Wang MH, Iclozan C, Fang B, et al. Sirt2 inhibition enhances metabolic fitness and effector functions of tumor-reactive T cells. Cell Metab. 2020;32:420-36.e12.
15. Ružić D, Đoković N, Nikolić K, Vujić Z. Medicinal chemistry of histone deacetylase inhibitors. Arh Farm. 2021;71:73-100.
16. Yang W, Chen W, Su H, Li R, Song C, Wang Z, et al. Recent advances in the development of histone deacylase SIRT2 inhibitors. RSC Adv. 2020;10:37382-90.
17. Sauve AA, Youn DY. Sirtuins: NAD+-dependent deacetylase mechanism and regulation. Curr Opin Chem Biol. 2012;16:535-43.
18. Wang Y, He J, Liao M, Hu M, Li W, Ouyang H, et al. An overview of sirtuins as potential therapeutic target: Structure, function and modulators. Eur J Med Chem. 2019;161:48-77.
19. Hong JY, Fernandez I, Anmangandla A, Lu X, Bai JJ, Lin H, et al. Pharmacological advantage of SIRT2-selective versus pan-SIRT1-3 inhibitors. ACS Chem Biol. 2021;16:1266-75.
20. Djokovic N, Ruzic D, Rahnasto-Rilla M, Srdic-Rajic T, Lahtela-Kakkonen M, Nikolic K, et al. Expanding the accessible chemical space of SIRT2 inhibitors through exploration of binding pocket dynamics. J Chem Inf Model. 2022;62:2571-85.
21. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021;19:4538-58.
22. Qian T, Zhu S, Hoshida Y. Use of big data in drug development for precision medicine: An update. Expert Rev Precis Med Drug Dev. 2019;4:189-200.
23. Zhu H. Big data and artificial intelligence modeling for drug discovery. Annu Rev Pharmacol Toxicol. 2020;60:573-89.
24. Roy K, Kar S, Das RN. A primer on QSAR/QSPR modeling. SpringerBriefs in Molecular Science. Cham: Springer International Publishing; 2015. ISBN 978-3-319-17280-4.
25. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: Where have you been? Where are you going to? J Med Chem. 2014;57:4977-5010.
26. Dixon SL, Duan J, Smith E, Von Bargen CD, Sherman W, Repasky MP, et al. AutoQSAR: An automated machine learning tool for best-practice quantitative structure-activity relationship modeling. Future Med Chem. 2016;8:1825-39.
27. Gramatica P. Principles of QSAR modeling: Comments and suggestions from personal experience. IJQSPR. 2020;5:61-97.
28. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47:D930-D940.

29. Chen L, Ai T, More S. Therapeutic compounds 2016. French Patent WO2016140978A1, 9 September 2016.
30. Swain MC, Cole JM. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model. 2016;56:1894-1904.
31. Landrum G. RDKit. Available from: http://rdkit.org (accessed 15 April 2022).
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. JAIR. 2002;16:321-57.
34. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: A molecular descriptor calculator. J Cheminform. 2018;10:4.
35. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv. 2016;arXiv:1603.04467.
36. O'Malley T, Bursztein E, Long J, Chollet F, Jin H, Invernizzi L, et al. Keras Tuner. 2019. Available from: https://keras.io/keras_tuner/ (accessed 30 April 2022).
37. Gramatica P. On the development and validation of QSAR models. Methods Mol Biol. 2013;930:499-526.
38. OECD. Validation of (Q)SAR models. Available from: https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed 20 August 2022).
39. Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model. 2009;49:1669-78.
40. Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R. External validation and prediction employing the predictive squared correlation coefficient: Test set activity mean vs training set activity mean. J Chem Inf Model. 2008;48:2140-5.
41. Ojha PK, Mitra I, Das RN, Roy K. Further exploring Rm2 metrics for validation of QSPR models. Chemom Intell Lab Syst. 2011;107:194-205.
42. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN. Some case studies on application of "Rm2" metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data. J Comput Chem. 2013;34:1071-82.
43. Chirico N, Gramatica P. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model. 2011;51:2320-35.
44. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. Molecules. 2012;17:4791-810.
45. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. J Med Chem. 2012;55:6582-94.
46. Riniker S, Landrum GA. Similarity maps—A visualization strategy for molecular fingerprints and machine-learning methods. J Cheminform. 2013;5:43.
47. Specs. Compound management services and research compounds for the life science industry. Available from: https://www.specs.net/ (accessed 8 January 2019).
48. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR, et al. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11:4.
49. Suvannang N, Preeyanon L, Ahmad Malik A, Schaduangrat N, Shoombuatong W, Worachartcheewan A, et al. Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study. RSC Adv. 2018;8:11344-56.
50. Zakharov AV, Zhao T, Nguyen DT, Peryea T, Sheils T, Yasgar A, et al. Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. J Chem Inf Model. 2019;59:4613-24.
51. Li S, Ding Y, Chen M, Chen Y, Kirchmair J, Zhu Z, et al. HDAC3i-Finder: A machine learning-based computational tool to screen for HDAC3 inhibitors. Mol Inform. 2021;40:e2000105.
52. Li R, Tian Y, Yang Z, Ji Y, Ding J, Yan A, et al. Classification models and SAR analysis on HDAC1 inhibitors using machine learning methods. Mol Divers. 2022.
53. Machado LA, Krempser E, Guimarães ACR. A machine learning-based virtual screening for natural compounds capable of inhibiting the HIV-1 integrase. Front Drug Discov. 2022;2:954911.

54. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods. 2000;44:235-49.
55. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des. 2007;13:3494-504.
56. Czub N, Pacławski A, Szlęk J, Mendyk A. Do AutoML-based QSAR models fulfill OECD principles for regulatory assessment? A 5-HT1A receptor case. Pharmaceutics. 2022;14:1415.
57. Roy K, Das RN, Ambure P, Aher RB. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intell Lab Syst. 2016;152:18-33.
58. Consonni V, Todeschini R, Ballabio D, Grisoni F. On the misleading use of Q2F3 for QSAR model comparison. Mol Inform. 2019;38:e1800029.
59. Chirico N, Gramatica P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model. 2012;52:2044-58.
60. Truchon JF, Bayly CI. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007;47:488-508.
61. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013. ISBN 978-1-4614-6848-6.
62. Li X, Kleinstreuer NC, Fourches D. Hierarchical quantitative structure-activity relationship modeling approach for integrating binary, multiclass, and regression models of acute oral systemic toxicity. Chem Res Toxicol. 2020;33:353-66.
63. Klingspohn W, Mathea M, ter Laak A, Heinrich N, Baumann K, et al. Efficiency of different measures for defining the applicability domain of classification models. J Cheminform. 2017;9:44.
64. Costantini S, Sharma A, Raucci R, Costantini M, Autiero I, Colonna G, et al. Genealogy of an ancient protein family: The sirtuins, a family of disordered members. BMC Evol Biol. 2013;13:60.
65. Lougiakis N, Gavriil ES, Kairis M, Sioupouli G, Lambrinidis G, Benaki D, et al. Design and synthesis of purine analogues as highly specific ligands for FcyB, a ubiquitous fungal nucleobase transporter. Bioorg Med Chem. 2016;24:5941-52.
66. Sklepari M, Lougiakis N, Papastathopoulos A, Pouli N, Marakos P, Myrianthopoulos V, et al. Synthesis, docking study and kinase inhibitory activity of a number of new substituted pyrazolo[3,4-c]pyridines. Chem Pharm Bull. 2017;65:66-81.
67. Blum CA, Ellis JL, Loh C, Ng PY, Perni RB, Stein RL, et al. SIRT1 modulation as a novel approach to the treatment of diseases of aging. J Med Chem. 2011;54:417-32.