RESEARCH ARTICLE

Enhancing cancer drug discovery: QSAR modeling with machine learning and chemical representations

Raúl Acosta-Murillo[1☯], José Carlos Ortiz-Bayliss[2☯], Patricio Adrian Zapata-Morin[1☯*]

1 Department of Microbiology and Immunology, School of Biological Sciences, Universidad Autónoma de Nuevo León, Pedro de Alba SN, San Nicolás de los Garza, Nuevo León, Mexico, **2** Tecnologico de Monterrey, School of Engineering and Sciences, Eugenio Garza Sada, Monterrey, Nuevo León, Mexico

☯ These authors contributed equally to this work.
* patricio.zapatamor@uanl.edu.mx

OPEN ACCESS

Citation: Acosta-Murillo R, Ortiz-Bayliss JC, Zapata-Morin PA (2026) Enhancing cancer drug discovery: QSAR modeling with machine learning and chemical representations. PLoS One 21(3): e0343654. https://doi.org/10.1371/journal.pone.0343654

Editor: Sunghwan Kim, National Library of Medicine, UNITED STATES OF AMERICA

Received: July 30, 2025

Accepted: February 8, 2026

Published: March 17, 2026

Data availability statement: The R2, MAE, RMSE of all the combination chemical representation language, model, dataset and the code used to train, test and save the models is available through GitHub https://github.com/RollerCoaster1899/AutoQSAR.

Abstract

Accurately predicting the bioactivity of small molecules against cancer therapeutic targets remains a significant challenge at the intersection of cheminformatics and drug discovery. This study comprehensively evaluates chemical representations, including AtomPair Counts (APC),Avalon (AVN), Extended-Connectivity Fingerprint diameter 4 (ECFP4), Extended-Connectivity Fingerprint diameter 6 (ECFP6), Feature-based Morgan 2 (FM2), Feature-based Morgan 3 (FM3), Mol2Vec (M2V), Molecular ACCess System (MACCS), Mordred 2D Chi Kappa (MK2), RDKFingerprint (RDF), Rdkit PhysChem (RDC), Torsion (TSN) combined with machine learning algorithms (Bayesian Ridge (BRG), Elastic Net (ENT), Extra Trees (ETT), Hist Gradient Boosting (HGT), K-Nearest Neighbors (*k*NN), Lasso (LSS), Multi-layer Perceptron (MLP), Partial least squares (PLS), Random Forest (RFT), Ridge (RDG), Support Vector Regressor (SVR), and XGBoost (XGB)) for predicting cancer bioactivities. The results show that while AVN chemical representation, in conjunction with SVR algorithm, achieved the highest predictive accuracy, with R^2 of 0.735 in FGFR1 dataset; The mTOR dataset demonstrated the highest average performance across all models and chemical representations, with an R^2 of 0.592 across various cancer datasets. These findings demonstrate how cheminformatics tools like molecular fingerprints and quantitative structure-activity relationship (QSAR) modeling can significantly enhance bioactivity prediction, ultimately contributing to more efficient and targeted cancer drug discovery.

Introduction

Overview of QSAR modeling and machine learning in drug discovery

QSAR models have been crucial in drug discovery, offering predictive insights into the biological activity and properties of untested compounds by predicting activity

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

based on molecular structure [1]. These models rely on various computational approaches to establish chemical structure and biological response relationships. However, selecting appropriate descriptors and modeling techniques remains a persistent challenge, with continuous innovation and debate regarding optimizing these models for improved accuracy and utility in drug discovery [2]. In recent years, machine learning (ML) techniques have increasingly been integrated with QSAR modeling to enhance predictive capabilities. This study evaluates the impact of these ML techniques and various molecular descriptors to identify combinations that optimize the predictive accuracy of QSAR models [3–5].

Challenges in predicting bioactivity for cancer therapeutics

Cancer remains one of the most significant global health challenges, with breast, prostate, and lung cancers among the most prevalent and deadly [6,7]. Accurate bioactivity prediction for small molecules targeting key cancer therapeutic proteins is essential for advancing cancer drug discovery. Recent research emphasizes the need to explore new therapeutic targets, such as HER2, PARP, PI3K, and mTOR, which play critical roles in cancer cell growth, proliferation, and metastasis [8]. Identifying molecules that effectively inhibit these targets requires sophisticated QSAR models and ML techniques capable of navigating the complexities of molecular structures and large datasets.

Molecular representations and descriptors in QSAR

Representing molecules in computational models poses several challenges, especially when dealing with non-canonical or non-unique molecular representations. To address this, a diverse array of molecular encoding schemes was employed. Substructure-based keys, including MACCS keys [9] and PubChem fingerprints, were used to offer binary resolutions of structural features. These were complemented by topological descriptors such as APC [10], TSN [11], and AVN [12], which describe substructural connectivity. Higher fidelity descriptors, specifically Extended Connectivity Fingerprints (ECFP4, ECFP6) and Feature-based Morgan fingerprints (FM2, FM3) [13], were applied to capture atomic environments within defined radii. These circular fingerprints are widely used in structure-activity modeling [14–17].

To capture semantic and contextual information providing new perspectives beyond traditional fingerprints, advances in continuous vector embeddings were leveraged. These included `M2V`, trained via the Word2Vec algorithm [18], and `SMILES-Vec`, which capture substructural relationships [19]. Furthermore, physicochemical descriptors play a crucial role in predicting bioactivity by offering insights into ADME (Absorption, Distribution, Metabolism, and Excretion) properties [20]. To this end, the RDKit descriptor suite and Mordred calculator were used to generate constitutional, electronic, and topological indices (e.g., Chi, Kappa, E-State) [21].

Dimensionality reduction techniques

Given the complexity and high dimensionality of molecular descriptors, effective feature selection is essential to simplify data analysis and enhance model

robustness. Instead of dimensionality reduction techniques like Principal Component Analysis (PCA), a stepwise feature filtering approach was employed to eliminate redundancy while preserving the original physicochemical meaning of the descriptors. First, invariant features were removed using a zero-variance filter (`VarianceThreshold`). Subsequently, a custom correlation filter was applied to identify and remove highly collinear features (Pearson correlation coefficient > 0.95) [22,23]. Importantly, to prevent data leakage, this feature selection process was fitted only on the training dataset prior to any cross-validation. The specific list of retained features identified in the training phase was then consistently applied to all held-out folds and external test sets. This strict separation ensures that the models are not biased by information from the test data, addressing common pitfalls in QSAR modeling such as overfitting and inflated performance estimates [24,25].

Machine learning models in QSAR

The integration of ML algorithms has fundamentally transformed QSAR modeling. While traditional linear approaches such as Ridge and Lasso regression established the foundation for interpreting physicochemical relationships [26,27], modern non-linear methods have significantly expanded predictive capabilities. Tree-based ensemble method, most notably Random Forest (RFT) and Gradient Boosting frameworks like XGBoost have emerged as some of the most effective tools, offering robust accuracy and computational efficiency in predicting bioactive properties [5,28,29]. Concurrently, kernel-based methods such as SVR and probabilistic approaches like Gaussian Processes have provided powerful alternatives for modeling complex non-linear landscapes [30,31]. More recently, deep neural networks (DNNs) have been applied to further push the boundaries of prediction, although their incremental improvements in accuracy compared to established ensemble models often come at the cost of increased computational complexity and reduced interpretability [32].

Recently, graph neural networks (e.g., GCNs, GATs, MPNNs, AttentiveFP) have enabled end-to-end molecular graph learning that often matches or surpasses RFT/XGB performance on large, multi-task QSAR datasets, albeit at higher GPU-dependent training costs [33,34]. Concurrently, transformer-based models such as ChemBERTa, the Molecule Attention Transformer, and MolE leverage self-attention on Simplified Molecular Input Line Entry System (SMILES) or graph embeddings to deliver state-of-the-art QSAR predictions after pre-training on millions of unlabeled molecules, though they demand substantial computational and memory resources [35,36]. Finally, interpretability methods like SHAP and attention-weight visualization decompose activity predictions into atom- and fragment-level contributions, bolstering medicinal chemists‘ confidence and guiding rational lead optimization [35,37].

Objectives and research questions

This work comprehensively evaluates machine learning models and molecular representations for predicting the bioactivity (pIC50) of small molecules targeting various cancer therapeutic proteins. We employ multiple sets of molecular descriptors to compare their performance in QSAR modeling. The main research question is: *Which combination of molecular representations and machine learning algorithms provides the highest predictive accuracy for bioactivity prediction of small molecules against cancer targets?*

Materials and methods

This section details the datasets, chemical representations, feature engineering techniques, machine learning models, and validation methods used in this study.

Data acquisition and preprocessing

Bioactivity data for sixteen therapeutic targets were obtained from the ChEMBL database [38], specifically focusing on IC_{50} measurements, which quantify the half-maximal inhibitory concentration of a compound. To ensure data consistency, we excluded records containing inequality relations (>, <), as well as records with missing values. This filtering step was

crucial to avoid introducing noise or uncertainty into the dataset. The molecular structures were standardized to their neutral, canonical SMILES representations using RDKit's Salt Remover utility, ensuring the removal of salts and other unimportant molecular components [39,40]. For compounds with multiple IC_{50} measurements, we calculated the median to represent a single, reliable activity value. We applied a ceiling of 10^8 nM to avoid distortion from exceedingly high IC_{50} values [41]. These IC_{50} values were subsequently converted to pIC_{50} values using the formula:

$$pIC50 = -\log 10(IC_{50} \times 10^{-9}) \quad (1)$$

Chemical representation

The core of the QSAR modeling process lies in the choice of molecular representations. In this study, we applied multiple molecular encoding schemes to capture diverse structural and physicochemical features of the compounds. These included molecular fingerprints such as ECFP4, ECFP6, FM2, FM3, and RDKIT Fingerprints [13], as well as topological fingerprints like MACCS [9], APC [10], TSN [11], and AVN [42], which describe substructural and connectivity information. Additionally, physicochemical descriptors derived from RDKit's built-in 2D descriptor suite were computed [40], alongside Mordred-calculated descriptors such as MK2, which describe molecular constitutional, electronic, and topological features [21].

To add more diversity to the molecular representations, we also used Mol2Vec embeddings. These are vector-based representations of molecules created using a Word2Vec model, which helps capture relationships between molecular substructures [18].

Feature engineering

Given the high dimensionality of the chemical representations, the datasets underwent feature engineering to eliminate irrelevant features. We applied the zero-variance filter, removing features with zero variance using `VarianceThreshold` [25]. Later, we used a custom-built correlation filter to remove highly correlated features, which can introduce multicollinearity and negatively impact model performance. The filter kept only the first feature in each highly correlated pair (with a Pearson correlation coefficient greater than 0.95), discarding the others. This step helped reduce redundancy and improve model interpretability [23]. Finally, we standardized the features using *Z*-score normalization (except for tree-based models, which are invariant to scaling), ensuring that all features had a comparable scale for models sensitive to feature magnitude, such as linear regression and support vector regressor (SVRs) [43].

Machine learning algorithms

In this study, we chose traditional machine learning methods for their interpretability and relatively low computational demands, aligning with our goal of establishing robust baseline models for molecular bioactivity prediction. Our methodology involved using various machine learning algorithms, including traditional models such as *k*NN [44] and PLS [45], alongside advanced techniques like SVR [46] RFT [47], and XGB [48]. Model training was conducted with the `Scikit-Learn` [25] and `XGBoost` [29] libraries.

While deep learning (DL) has shown potential in QSAR applications, recent studies, such as the IDG-DREAM Drug-Kinase Binding Challenge (2019), suggest that DL's predictive improvement over traditional methods like XGB or RFT can be modest [5]. This limited performance advantage, combined with the high computational demands and interpretability challenges associated with DL, led us to focus on traditional approaches in this phase of our study.

***k*-nearest neighbors (*k*NN):** A non-parametric method that predicts outcomes based on proximity to the *k* closest points in the training data [49]. The number of neighbors was optimized over the set $k \in \{3, 5, 7, 11, 15\}$.

Partial least squares (PLS): A technique that handles multicollinearity by extracting latent factors that maximize the covariance between independent and dependent variables [50]. The model was tuned by selecting the optimal number of latent components from $\{2, 5, 10\}$.

Support vector regressor (SVRs): SVRs create a hyperplane that maximizes the margin or minimizes regression errors using kernel functions [51].We utilized the Radial Basis Function (RBF) kernel, optimizing the regularization parameter C (log-scale 10^{-2} to 10^{3}) and the kernel coefficient γ (log-scale 10^{-4} to 10^{0}).

Random Forest (RFT): RFT creates an ensemble of decision trees with random feature selection [52]. Hyperparameters tuned included the number of estimators ($\{100, 300\}$) and the maximum features considered for splitting ($\sqrt{n\_features}$ or $\log_2 n\_features$).

Extreme Gradient Boosting (XGB): XGB sequentially builds weak learners to reduce residual errors [29]. The grid search included the number of estimators ($\{200, 500\}$), learning rate ($\{0.01, 0.05, 0.1\}$), and maximum tree depth ($\{3, 6, 9\}$).

ElasticNet (ENT): A linear model combining L1 and L2 regularization [53]. Optimization was performed on the regularization strength α (10^{-4} to 10^{2}) and the L1 ratio ($\{0.1, 0.5, 0.9\}$) to balance Lasso and Ridge penalties.

ExtraTrees (ETT): An ensemble method using extremely randomized trees [54]. Similar to RF, we optimized the number of estimators ($\{100, 300\}$) and the feature subset size for splitting ($\sqrt{n\_features}$ or $\log_2 n\_features$).

HistGBDT (HGT): A histogram-based Gradient Boosting method that bins continuous features for speed [55]. We tuned the maximum number of iterations ($\{200, 500\}$) and the learning rate ($\{0.01, 0.05, 0.1, 0.2\}$).

Lasso (LSS): A regression analysis method that performs variable selection using L1 penalties [27]. The regularization strength α was optimized on a logarithmic scale.

Multi-layer Perceptron (MLP): A feedforward artificial neural network trained via backpropagation [56]. The architecture search included single hidden layers of 50 or 100 neurons and a two-layer configuration $(100, 50)$, with L2 regularization $\alpha \in \{10^{-4}, 10^{-2}\}$.

Ridge (RDG): A linear regression model with L2 regularization [26]. The regularization strength α was optimized over a logarithmic scale ranging from 10^{-4} to 10^{4}.

Bayesian Ridge (BRG): A probabilistic linear model that estimates coefficients with Gaussian priors [57]. This model self-tunes regularization parameters during fitting without an external grid search.

Training and hyperparameter optimization

The pipeline supports 12 distinct machine learning models, spanning a variety of algorithmic paradigms. These include linear models such as Ridge, Lasso [27], ENT, and BRG; tree-based ensemble methods such as RFT [28], ETT, HGT, XGB [29], SVR [30]; and other varied methods such as *k*-NN [58], and multi-layer perceptron neural networks [56]. We conducted hyperparameter optimization for each model using Halving Randomized Search Cross-Validation, which efficiently narrows the search space by iteratively halving the number of candidate configurations based on performance [59]. We employed a group three-fold cross-validation scheme with Murcko scaffolds [60] as grouping variables, to ensure that training and validation sets did not overlap in scaffold composition, preventing data leakage and reducing the risk of overfitting [41]. We tested 20 candidate hyperparameter sets per model, and identified the best-performing configurations. Upon finding optimal configurations, we retrained each model on the entire training set. We evaluated them on an unseen test set comprising approximately 15% of the molecules to ensure unbiased performance assessment [61–63].

Model evaluation and statistical analysis

To evaluate model performance, we used the coefficient of determination (R^2) as the primary metric, computed on an external test set generated using Murcko scaffold splitting to ensure that test compounds possessed distinct molecular frameworks from the training set [60]. Outliers were identified and removed from the predicted registry using the 1.35 Interquartile Range (IQR) technique, ensuring that extreme predictions did not distort the results [64]. In addition to R^2,

we calculated RMSE and Mean Absolute Error (MAE) to provide a comprehensive view of prediction accuracy and error magnitude. The distribution of R^2 scores was analyzed using the Friedman test to identify global performance differences between models and representations [65]. We conducted this using the `scipy.stats.friedmanchisquare` function [66]. Where significant differences were identified (p-value < 0.05), we conducted the Nemenyi post-hoc test to assess pairwise differences, visualizing the resulting p-values via heatmaps. Additionally, pairwise T-tests were conducted to determine statistical significance between specific model and representation pairs. To investigate model stability and consistency, we performed a ranking analysis based on median R^2 scores across datasets. We complemented this by analyzing the trade-off between performance and variability, computing the standard deviation of R^2 (model disagreement) and determining dataset difficulty, defined as ($1 - \text{Best } R^2$) based on Cohen's framework [67]. These metrics allowed us to distinguish between stable predictive power and high variability across the different cancer therapeutic targets.

Results

To evaluate the influence of molecular representations, machine learning algorithms, and preprocessing strategies, we selected 16 biological targets associated with breast, prostate, and lung cancers [8,68]. The bioactivity data for these targets were curated and encoded using 12 types of molecular representations [16,19]. Data preprocessing included removing low-variance features and eliminating multicollinearity. Furthermore, supervised feature selection was applied using univariate statistical tests and Lasso-based importance. We used these processed features to train 12 machine-learning QSAR models [5], with hyperparameters optimized via a randomized search strategy. This workflow resulted in a comprehensive set of trained models, as illustrated in Fig 1. Finally, we conducted statistical tests to assess significant differences in predictive performance.

After scouring the ChEMBL database, we identified 16 distinct cancer biological targets, each accompanied by its ChEMBL ID, UniProt ID, gene, gene description, and the cancer they are related to, as presented in Table 1. Additionally, we provide the number of IC50 entries for each target. The ChEMBL datasets vary in size, ranging from 2,444 entries for the CHK1 protein to 16,715 entries for the EGFR1 receptor. On average, the datasets contain 5,625 entries.

Impact of the chemical representation on QSAR performance

To investigate how different chemical representations impact QSAR model performance, we evaluated a range of molecular descriptors. These included various molecular fingerprints, such as ECFP4, ECFP6, FM2, FM3, and RDKit Fingerprints [13], as well as topological fingerprints like MACCS [9], APC [10], TSN [11], and AVN [42]. We also assessed RDKit descriptors [40] and Mordred-calculated descriptors [21]. The average R^2 values for each chemical representation across all datasets and models are displayed in Table 2 and Fig 2, while Fig 3 presents the distribution of the MAE. Additionally, Fig 4 presents a heatmap of R^2 values, providing an overview of average performance across both representation languages and model architectures (averaged across datasets). While In the Fig 5 it can be observed the best R^2 fount across the datasets.

While the best R^2 was observed in combination of AVN and SVR (0.591); In average, the RDF molecular fingerprint emerged as the top performer, achieving the highest average R^2 value of 0.510 across datasets, and model architectures. It was closely followed by AVN (0.489) and ECFP6 (0.460). This ranking is also illustrated in Fig 6 and supported by the stability analysis in Fig 7, where the average ranks for the representations are as follows: RDF (2.13), AVN (3.20), and ECFP6 (5.07). This suggests that RDF was the most effective representation for capturing the relationship between chemical structures and biological activity in our study. Statistically significant differences were observed, with the exception of the comparison between RDF (1st) and AVN (2nd), as detailed in Figs 8 and 9.

Impact of the machine learning model on QSAR performance

To further explore the role of machine learning models in QSAR performance, we tested a suite of 12 algorithms trained for pIC50 prediction: BRG, ENT, ETT, HGT, *k*NN, LSS, MLP, PLS, RFT, RDG, SVR, and XGB (see Methods

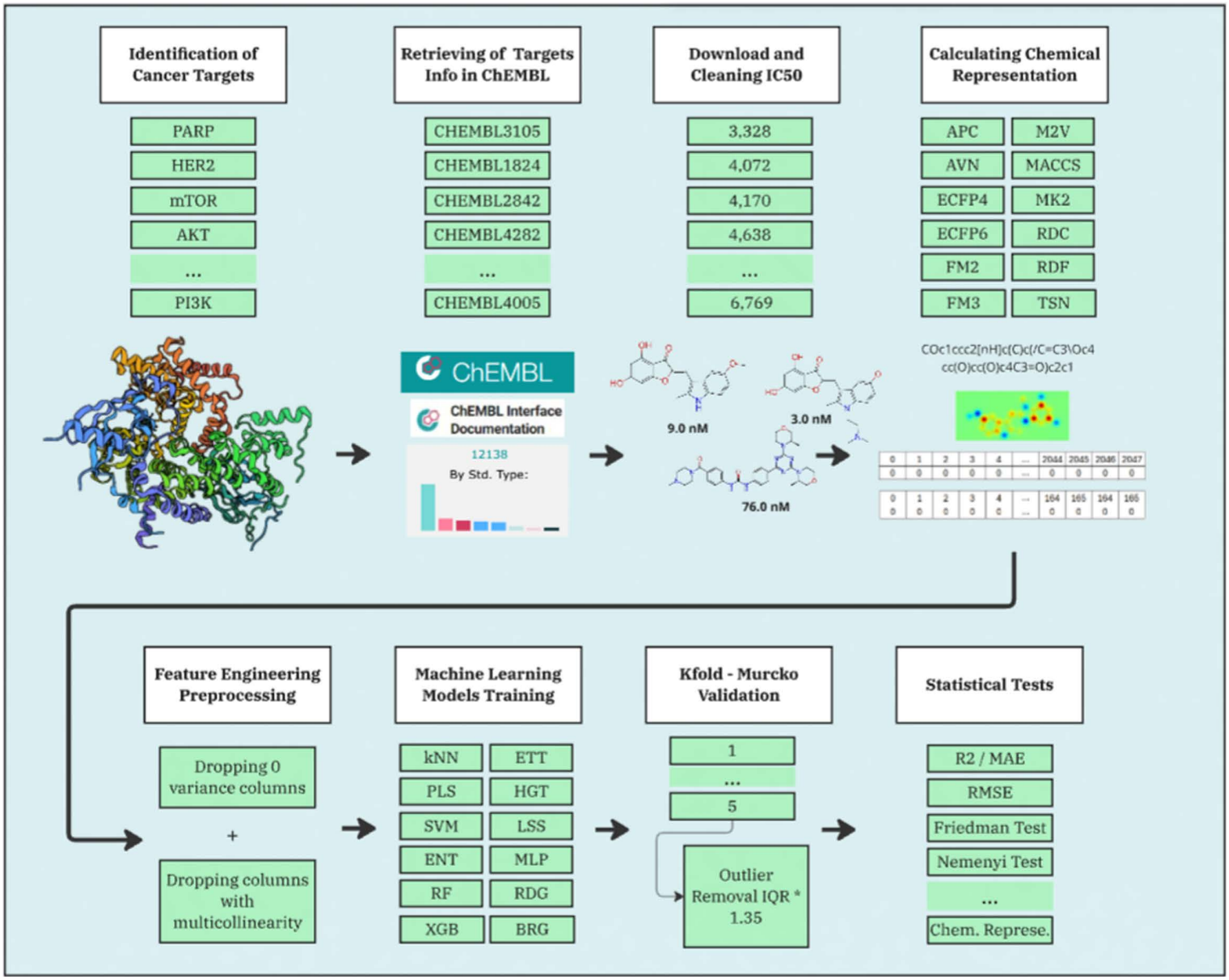

Fig 1. Workflow diagram of the QSAR modeling pipeline incorporating chemical representations and machine learning algorithms. Methodological process diagram detailing the research workflow, including dataset selection, chemical representations (such as APC, AVN, ECFP4, ECFP6, FM2, FM3, M2V, MACCS, MK2, RDF, RDC, TSN), and machine learning algorithms applied (BRG, ENT, ETT, HGT, *k*NN, LSS, MLP, PLS, RFT, RDG, SVR, XGB). The results were validated, ensuring robust model performance evaluation across all tested combinations.

https://doi.org/10.1371/journal.pone.0343654.g001

for additional details). The average R^2 and RMSE values for each model across all datasets and chemical representations are provided in Figs 3 and 10, while we show the distribution of MAE values in Fig 11. As in the previous analysis, Fig 4 visualizes the heatmap of R^2 values, again averaged across the datasets. In the Fig 5 it is shown the best R^2 fount across the datasets.

Among the models tested, ETT outperformed all others with the highest average R^2 of 0.541, followed by HGT and RFT, both with an R^2 of 0.530. This ranking is further corroborated in Fig 12 and validated by the stability analysis in Fig 13, where the average ranks for the models are: ETT (1.53), HGT (2.60), and RFT (2.2.93). Nemenyi plot between the models

Table 1. Description of the identified cancer targets with ChEMBL and UniProt IDs.

ChEMBL ID	Gene	UniProt ID	Description	# IC50	Related
CHEMBL3105	PARP	P09874	Poly [ADP-ribose] polymerase-1	3,328	B, L, T
CHEMBL1824	HER2	P04626	Receptor protein-tyrosine kinase	4,072	B, L
CHEMBL4005	PI3K	P42336	PI3-kinase p110-alpha subunit	6,769	B, L, T
CHEMBL3130	PI3K	O00329	PI3-kinase p110-delta subunit	4,638	B, L, T
CHEMBL4282	AKT	P31749	Serine/threonine-protein kinase AKT	4,170	B, L, T
CHEMBL2842	mTOR	P42345	Serine/threonine-protein kinase	5,112	B, L, T
CHEMBL3650	FGFR1	P11362	Fibroblast growth factor receptor 1	4,820	B, L
CHEMBL2742	FGFR3	P22607	Fibroblast growth factor receptor 3	3,375	B, L
CHEMBL1871	AR	P10275	Androgen Receptor	3,670	B, L
CHEMBL203	EGFR1	P00533	Epidermal growth factor receptor	16,715	B, L
CHEMBL1957	IGFR1	P08069	Insulin-like growth factor I receptor	4,424	B, L, T
CHEMBL4630	CHK1	O14757	Serine/threonine-protein kinase Chk1	2,444	B, L, T
CHEMBL279	VEGFR2	P35968	Vascular endothelial growth factor receptor 2	14,087	B, L, T
CHEMBL4722	AURKA	O14965	Serine/threonine-protein kinase Aurora-A	3,725	B, L, T
CHEMBL2185	AURKB	Q96GD4	Serine/threonine-protein kinase Aurora-B	2,367	B, L, T
CHEMBL1865	HDAC6	Q9UBN7	Histone deacetylase 6	6,287	B, L, T

https://doi.org/10.1371/journal.pone.0343654.t001

Table 2. Comparison of R^2 and RMSE Performance for pIC50 Across Different Chemical Representations. Values represent the coefficient of determination (R^2) and RMSE. The highest-performing representation (highest R^2 and lowest RMSE) is indicated in bold.

Representation	R^2	RMSE
APC	0.3877	0.8495
AVN	0.4887	0.7749
ECFP4	0.4522	0.7902
ECFP6	0.4596	0.7802
FM2	0.4441	0.7981
FM3	0.4523	0.7927
M2V	0.4092	0.8373
MACCS	0.4041	0.8444
MK2	0.3925	0.8456
RDC	0.4162	0.8303
RDF	**0.5101**	**0.7466**
TSN	0.4538	0.7951

https://doi.org/10.1371/journal.pone.0343654.t002

can be obseved at Fig 14 Statistically significant differences were found, except between ETT (1st), HGT and SVR, as shown in Fig 15.

Impact of datasets on QSAR performance

In addition to model performance, we also evaluated 16 different datasets for pIC50 prediction. Fig 16 again features a heatmap of R^2 values, averaged across models and representations, while Fig 17 shows the distribution of MAE. The Table 4 and Fig 18 present the average R^2 and RMSE values across datasets, models, and chemical representations.

Building upon Cohen's framework [67], we define the *difficulty* of a dataset as (1 – Best R^2). The difficulty scores for the various datasets are presented in Fig 19. In addition, Fig 20 illustrates the standard deviation of R^2, revealing that

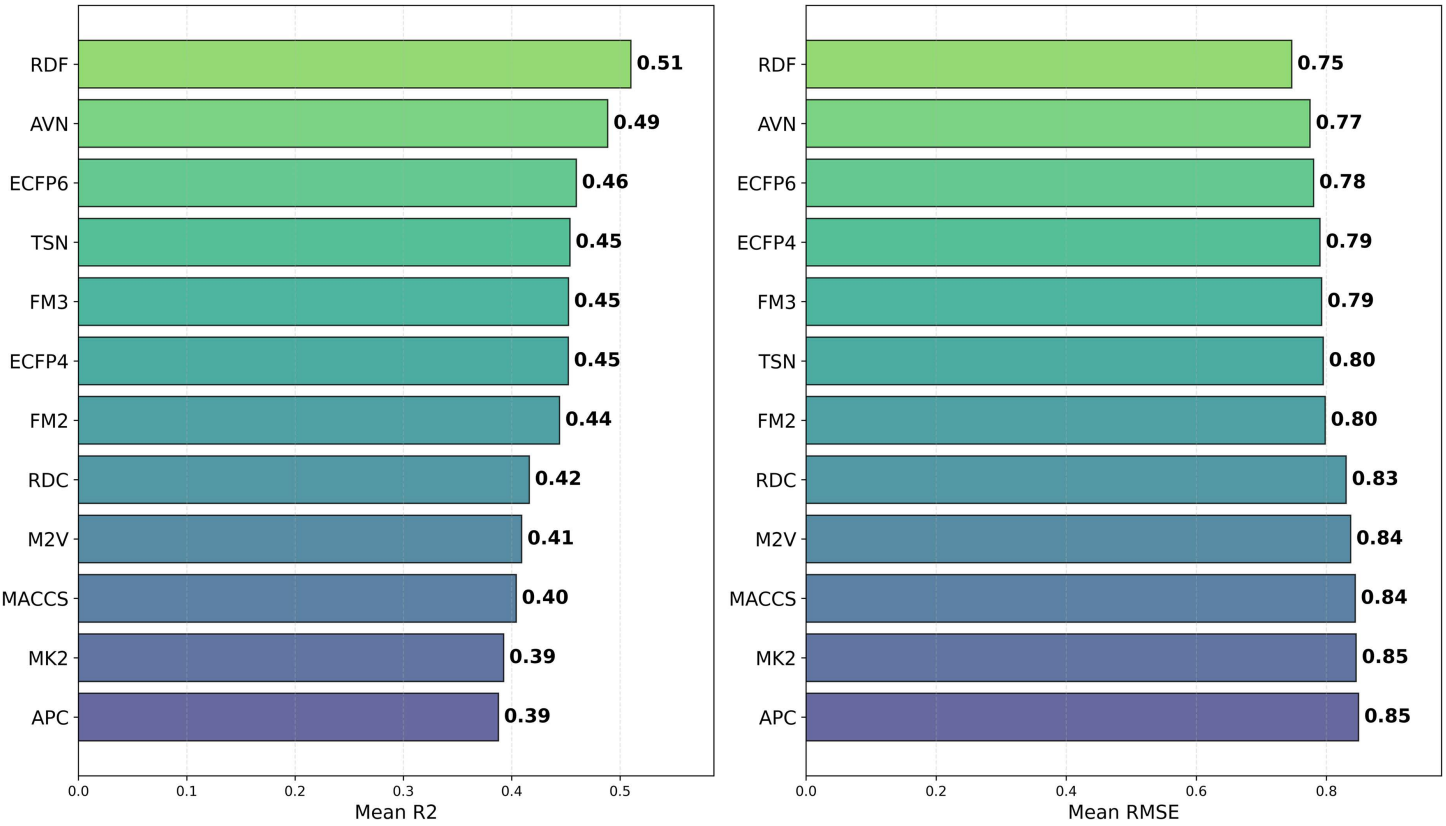

Fig 2. Distribution of R^2 and RMSE for pIC50 Across Chemical Representations. Barplot (left) showing average R^2 and (right) MAE of prediction values for pIC50 across multiple cancer therapeutic datasets and model architectures, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g002

the model with the highest average R^2 did not necessarily exhibit the smallest model disagreement. This observation suggests that a higher R^2 does not always imply reduced variability in model predictions, underscoring the complexity of model performance. This is further emphasized in Fig 21, where datasets with lower difficulty (on the left side) show model disagreements in the range of 0.1 to 0.2, highlighting how even less challenging datasets can exhibit considerable model variability.

Among the datasets, mTOR demonstrated the highest average R^2 of 0.592, followed by HER2 (0.586) and CHK1 (0.582), as seen in Fig 18. These results suggest that mTOR was the dataset most strongly correlated with biological activity, with a higher R^2 indicating more reliable predictive power based on molecular structure [24]. This is also supported by the ranking in Fig 22 and the stability analysis in Fig 23, where the average ranks for the datasets are: mTOR (1.92), HER2 (2.08), CHK1 (2.42). Nemenyi values between datasets can be observed in Fig 24. Statistically significant differences were observed, except between mTOR (1st), PARP, and PI3K, as shown in Fig 25.

Discussion

We have analyzed the impact of various chemical representations, including MACCS keys [9], PubChem fingerprints, APC [10], TSN [11], AVN [12], ECFP4, ECFP6, Feature-based Morgan fingerprints (FM2, FM3) [13], M2V [18], RDKit descriptor suite and Mordred calculator [21]. These chemical representations are characterized by their widespread adoption and efficiency. According to our experiments, RDF obtained the highest average R^2 values among all the chemical

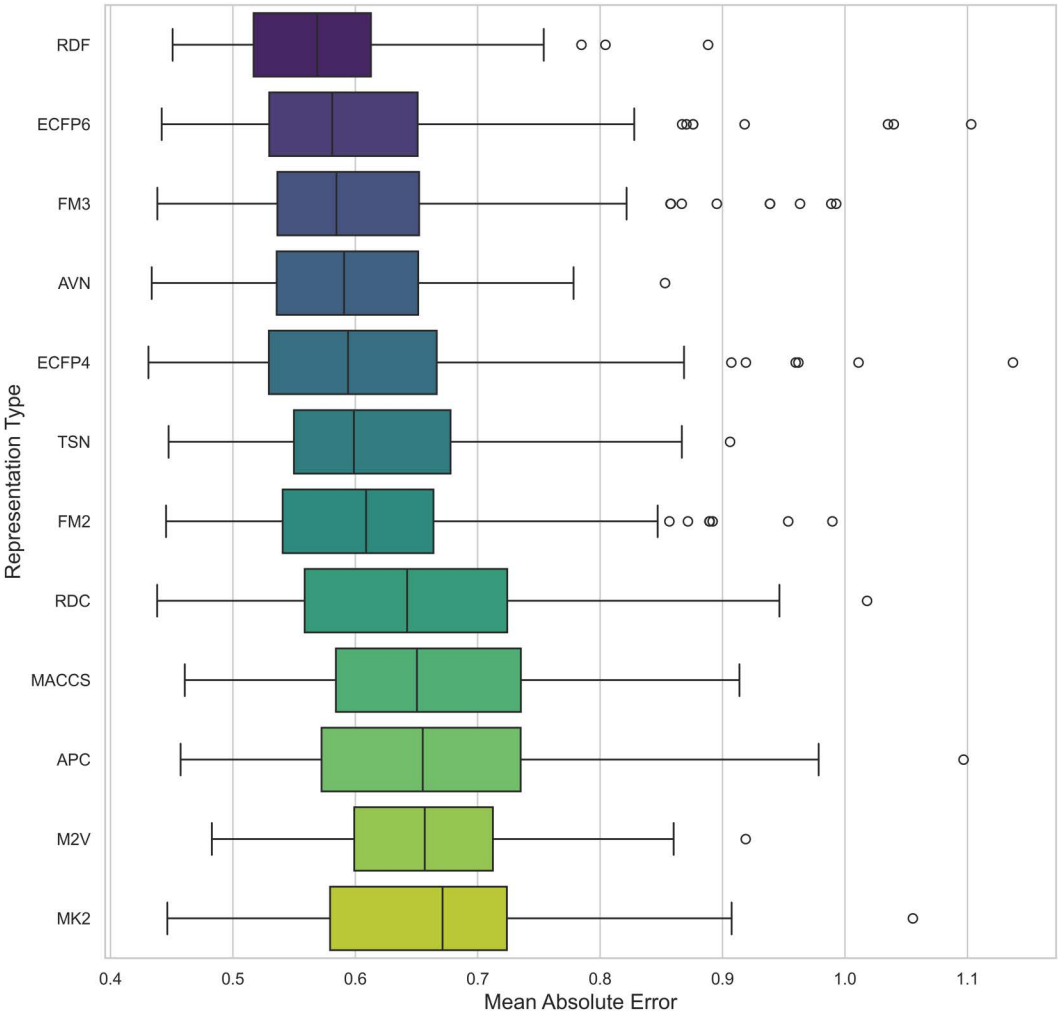

Fig 3. Distribution of MAE for pIC50 Across Chemical Representations. Boxplot showing the average MAE of prediction values for pIC50 across multiple cancer therapeutic datasets, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g003

representations (0.510), surpassing AVN (0.489) and ECFP6 (0.460).All the other molecular descriptors exhibited lower performances, with APC being the lowest (0.388), with an R^2 difference of 0.122 compared to AVN, as shown in Table 2 and Fig 2. Such differences were statistically significant, except when comparing RDF with AVN ($p = 0.09$) (2nd place) (see Table 8 for details).

Our findings are consistent with those of Sabando et al. [19], who compared molecular representations based on word embeddings with traditional approaches, such as MACCS and ECFP fingerprints, in both regression and classification tasks. Their study found no significant performance improvements when embedding methods were applied to QSAR modeling. Similarly, in our analysis, we observed an average ranking of 8.5 for the M2V embedding model, while RDF and AVN models achieved rankings of 2.06 and 3.25, respectively. Similarly, Orosz et al. [69] reported comparable predictive performance between MACCS and ECFP4 for pharmacodynamics and toxicity predictions using an XGB model. In our analysis, ECFP4 yielded an average R^2 of 0.452, while MACCS produced a value of 0.404. However, the difference between these models was not statistically significant ($p = 0.06$) according to the t-test (see Table 8). Lee et al. [70] further

Fig 4. Distribution of R^2 Prediction Values for pIC50 Across Chemical Representations and Model Architectures. Heatmap showing the average R^2 prediction values for pIC50 across multiple cancer therapeutic datasets, organized by chemical representations and model architectures.

https://doi.org/10.1371/journal.pone.0343654.g004

supported this notion by reporting similar average accuracies between MACCS and ECFP4 for biodegradation prediction. Xie et al. [71] demonstrated that combining ECFP4 fingerprints with MACCS keys can enhance the accuracy of predicting the LogP parameter, highlighting the complementary nature of these representations. However, as noted by Zagidullin et al. [72], model performance should also be evaluated using more qualitative considerations, such as downstream tasks and model characteristics, to fully capture and enhance the capabilities of these approaches.

The superior performance of RDF descriptors compared to substructural fingerprints like ECFP and MACCS suggests that predicting cancer bioactivity requires capturing more than just local substructural fragments. While binary fingerprints encode the presence or absence of specific groups, RDF descriptors capture the probability distribution of atomic properties at varying radial distances, effectively encoding steric constraints and inter-atomic relationships in a continuous manner [73]. This aligns with the findings of highly specific protein-ligand interactions, where the spatial arrangement of atoms, mimicking the pharmacophoric geometry of the binding pocket is often more predictive than topological connectivity alone [74]. By retaining information about the global topology and continuous electronic distribution, RDF descriptors likely avoid the “information loss” inherent in the bit-collision hashing processes of standard circular fingerprints.

The underwhelming performance of continuous embeddings like M2V relative to explicit descriptors suggests that unsupervised pre-training on general chemical space does not always translate to specific bioactivity endpoints. Continuous vector representations tend to smooth the feature space, which can obscure “activity cliffs” instances where a minor

Fig 5. Distribution of R^2 Prediction Values for pIC50 Across Chemical Representations and Model Architectures. Heatmap showing the maximum R^2 prediction values for pIC50 across multiple cancer therapeutic datasets, organized by chemical representations and model architectures.

https://doi.org/10.1371/journal.pone.0343654.g005

structural modification leads to a disproportionate change in potency [75]. Discrete descriptors and fingerprints are often better equipped to flag these specific structural alerts. Furthermore, learned representations typically require substantially larger training sets to fine-tune the embeddings for specific downstream tasks, suggesting that our dataset sizes were insufficient to leverage the full semantic power of the M2V architecture, as highlighted by Winter et al. [76].

From a practical standpoint, the trade-off between computational cost and predictive accuracy is a critical consideration for virtual screening campaigns. While RDF descriptors provided statistical improvements, they are computationally more intensive to generate than bit-vector fingerprints like ECFP4. Riniker and Landrum [42] demonstrated that circular fingerprints are among the fastest to compute, making them ideal for initial high-throughput screening of ultra-large libraries. Therefore, while RDF coupled with ETT offers the highest precision for lead optimization, a streamlined pipeline using ECFP-based XGB models may be more efficient for the early-stage filtering of millions of compounds, providing a balance between speed and acceptable predictive power [77].

Here, we evaluated the impact of some commonly used machine learning algorithms to generate models for QSAR. The algorithms, ranked by their R^2 values, are ETT, HGT,RFT, XGB, SVR, etc., being ETT the best performer, as shown in Table 3 and Fig 12. The values for ETT did not present statistically significant differences from HGT (2nd place), RFT (3rd) (which can be observed at Fig 12). At the same time, MLP and ENT showed the worst performance, with an R^2 of 0.204 and 0.381, respectively. The preference for the RFT model in QSAR is due to its high predictive level and low number of

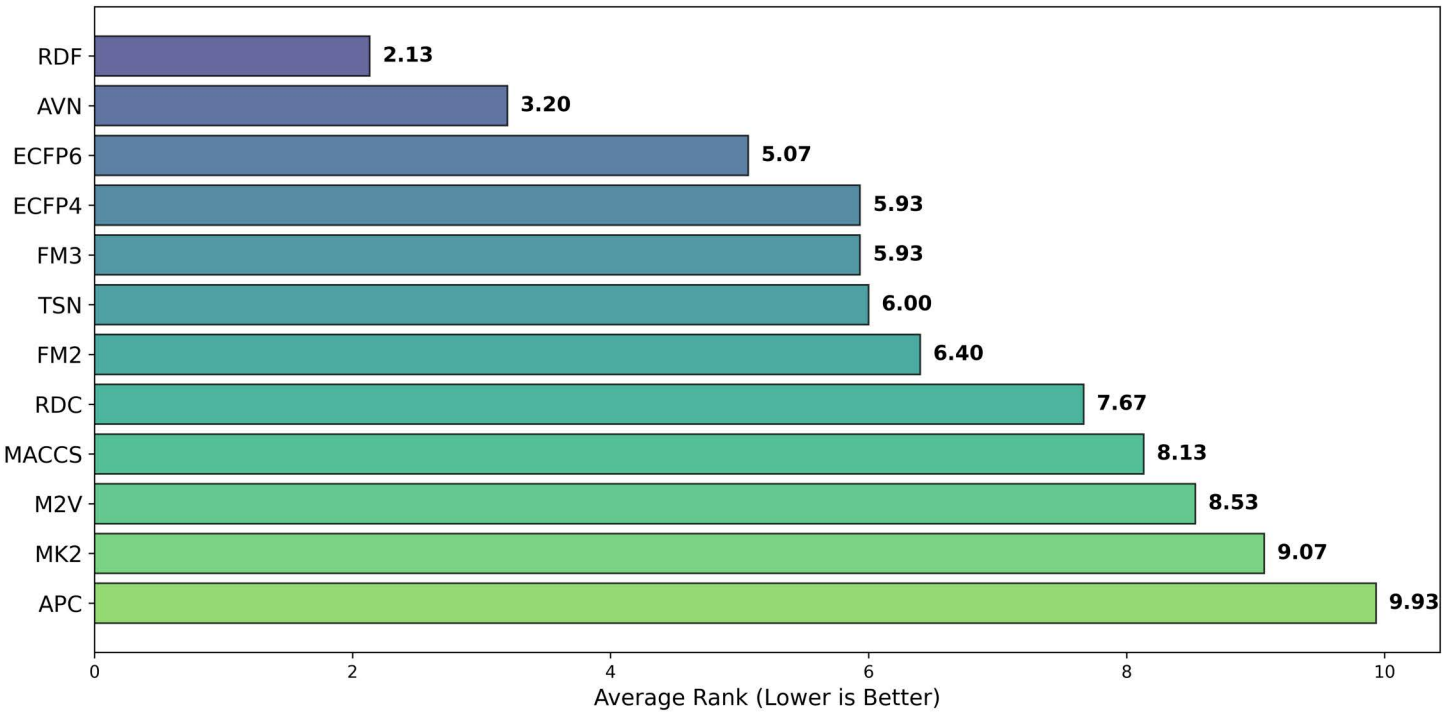

Fig 6. Average Ranking of R^2 for pIC50 Across Chemical Representations. Barplot showing the average ranking based on R^2 for pIC50 across multiple cancer therapeutic datasets and model architectures, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g006

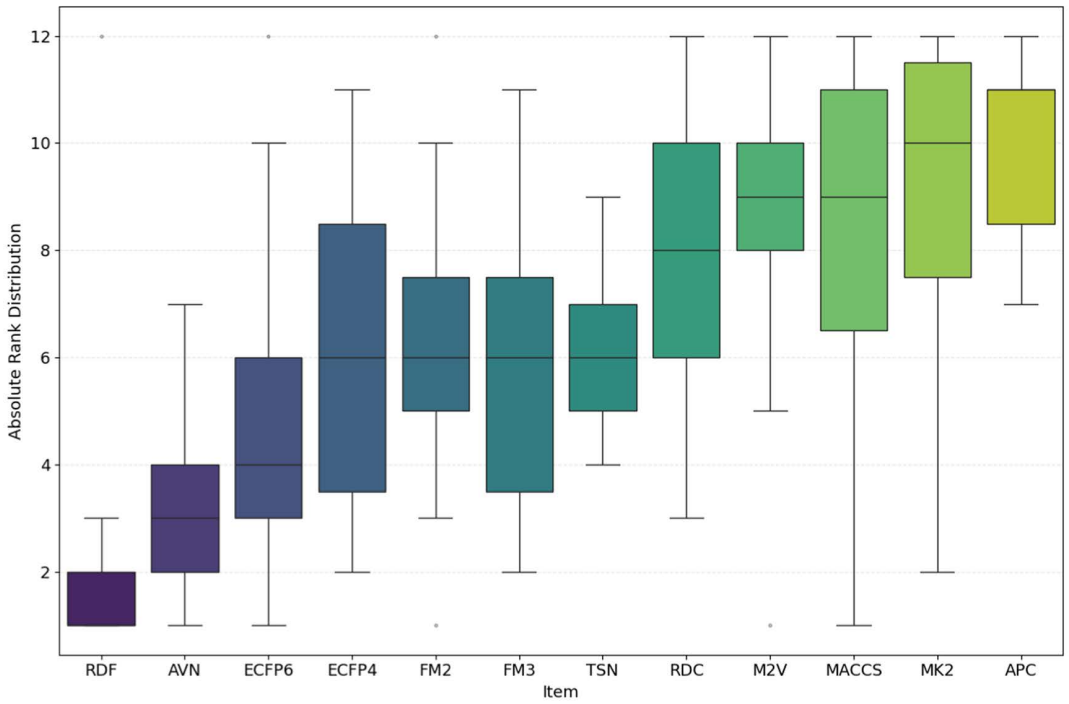

Fig 7. Ranking Distribution of R^2 for pIC50 Across Chemical Representations. Boxplot showing the ranking based on R^2 for pIC50 across multiple cancer therapeutic datasets and model architectures, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g007

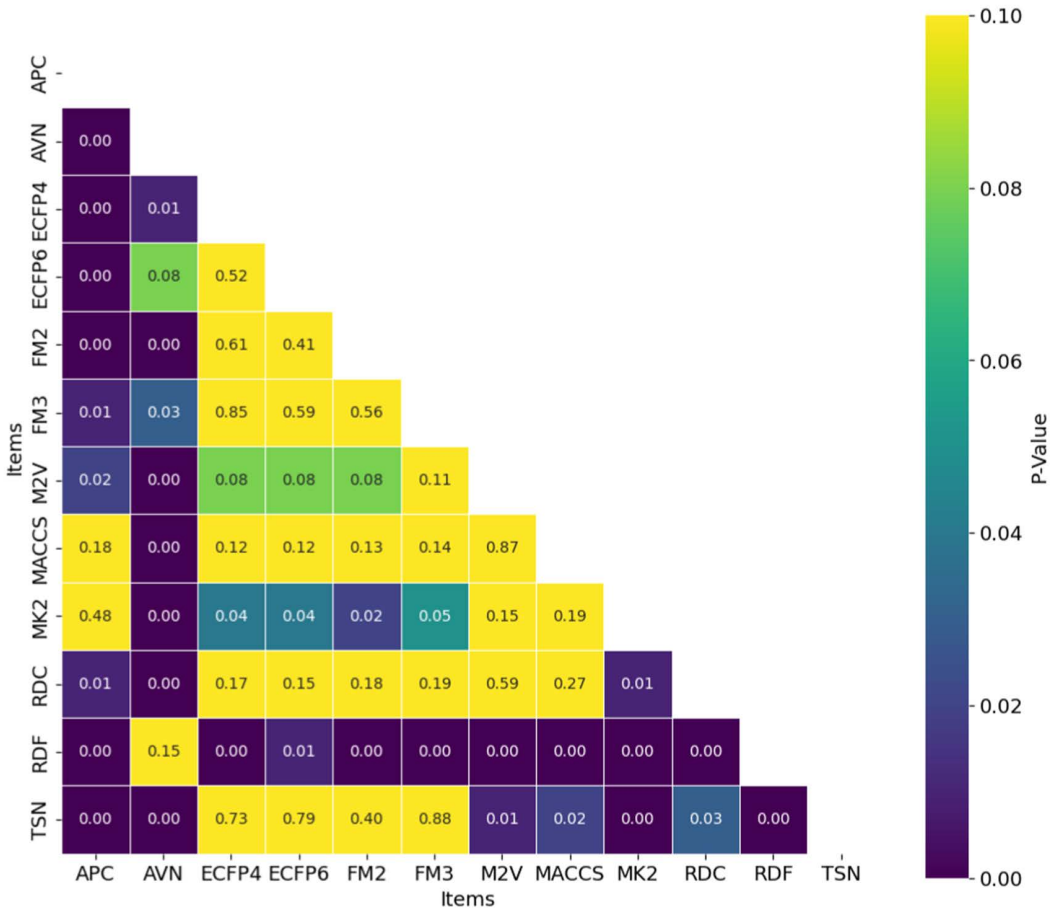

Fig 8. Distribution of T-test *p*-values for Statistical Significance of R^2 Across Chemical Representations and Model Architectures. Heatmap showing T-test *p*-values for statistical significance of R^2 across multiple cancer therapeutic datasets, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g008

adjustable parameters. At the same time, the XGB model is also recommended for its high accuracy and speed, where, regardless of the representation language, the RFT model has an average R^2 value similar to XGB (0.530 and 0.521, respectively). The most effective combination was observed when usingAVN with SVR, as well as when using ECFP6 and ECFP4 with ETT models in Fig 4.

The role of the ML model on the bioactivity prediction is highlighted by the works of Du et al. [78] and Wiriyarattanakul et al. [79]. In the former, different machine learning methods were evaluated for predicting the antioxidant activity of tripeptides, with *k*NN models outperforming XGB, RFT, and SVR in terms of accuracy (obtaining 0.996, 0.987, 0.945, and 0.926, respectively). Similarly, the latter conducted comparative studies for the QSAR prediction of anti-inflammatory activity, revealing a descending accuracy trend of SVM, GBR, and RFT, with accuracies of 0.907, 0.806, and 0.724, respectively. Similar results were found in inhibitor classification studies, where RFT demonstrated superior performance with an acurracy of 0.91 compared to PLS, which had an acurracy of 0.69 [80]. Additionally, for antibacterial compound classification, random forest and kNN achieved the highest accuracy (0.97) [81]. Lane et al. evaluated over 5,000 datasets using various algorithms for classification tasks, reporting SVC (0.796) and RFT (0.795) as the methods with better ROC-AUC, outperforming deep learning methods and kNN [82].

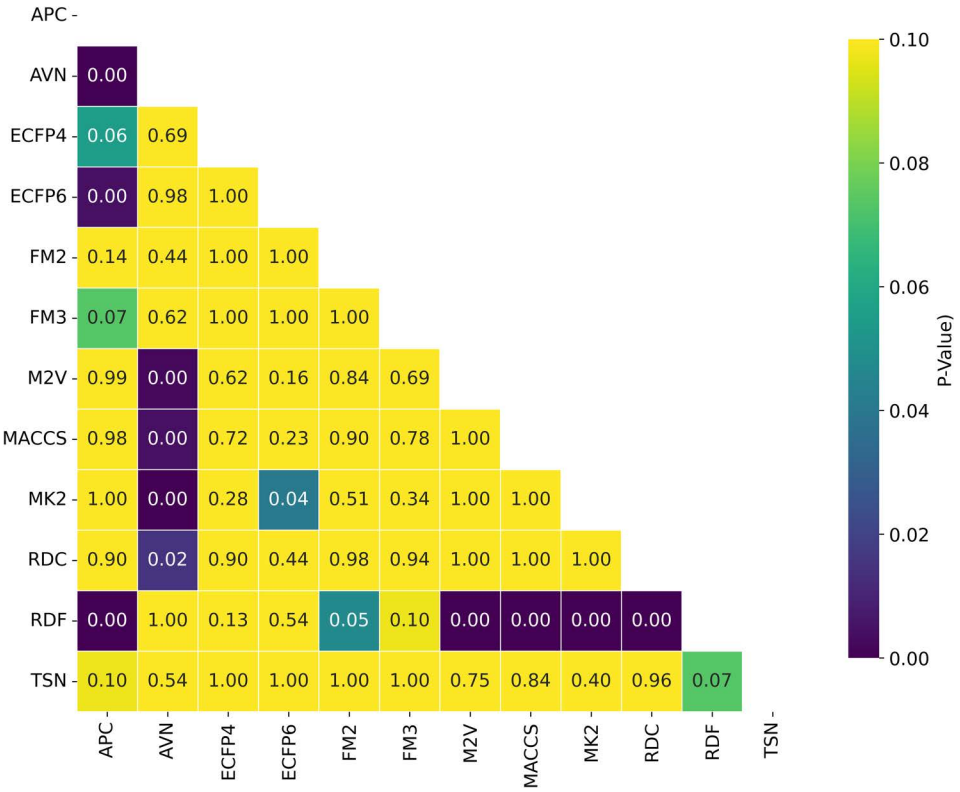

Fig 9. Distribution of Nemenyi p-values for Ranking Statistical Significance Across Chemical Representations and Model Architectures. Heatmap showing Nemenyi *p*-values for ranking statistical significance across multiple cancer therapeutic datasets, organized by chemical representations.

https://doi.org/10.1371/journal.pone.0343654.g009

Our findings reinforce a growing consensus in the cheminformatics community: for tabular QSAR datasets of moderate size (N less than 100k), tree-based ensemble methods often outperform standard deep neural networks. The underperformance of the MLP model in this study (Average R^2 = 0.202)) can be attributed to the lack of inductive bias suitable for tabular chemical data and the tendency of neural networks to overfit on smaller datasets without extensive pre-training [83]. Conversely, ensemble methods like ExtraTrees and XGBoost effectively handle high-dimensional feature spaces and reduce variance through bagging and randomization. As noted by Wu et al. [41], While deep learning dominates big data applications, traditional algorithms retain a distinct advantage in robustness and efficiency when applied to the limited datasets typical of bioactivity prediction.

Expanding upon our results, to assess the impact of the dataset on the R^2 metric of the models, records of 15 therapeutic targets with more than 2,000 entries for breast, prostate, and lung cancer were obtained from the ChEMBL platform (Table 1). Crucially, the testing sets were generated using Murcko scaffold splitting [60]. Unlike random splitting, this approach ensures that the test compounds possess distinct molecular frameworks from the training set, thereby rigorously evaluating the models' ability to generalize to new chemical spaces. Subsequently, we trained different machine learning models, which we later evaluated using various metrics. mTOR showed the highest R^2 value (0.592), while AR had the worst performance (0.227), as seen in Table 4 and Fig 18.

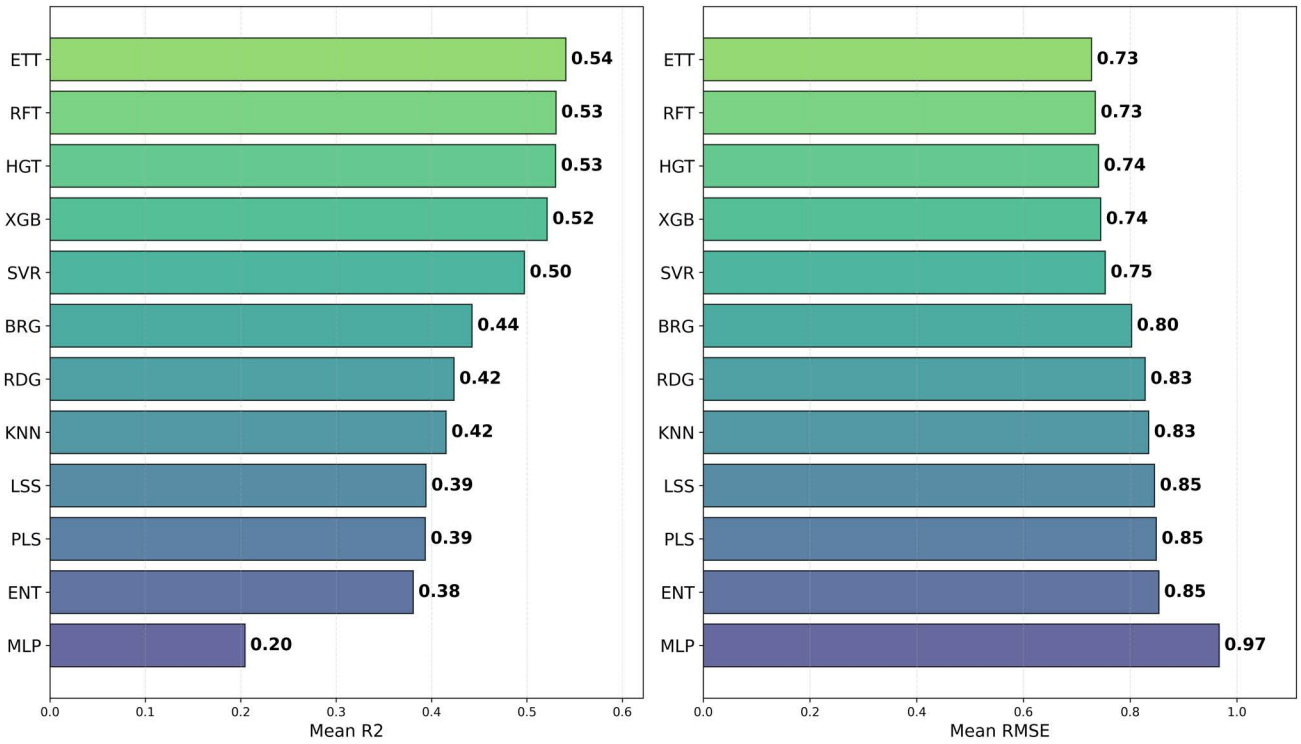

Fig 10. Distribution of R^2 and RMSE for pIC50 Across Model Architectures. Boxplot (left) showing average R^2 and (right) MAE of prediction values for pIC50 across multiple cancer therapeutic datasets and chemical representations, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g010

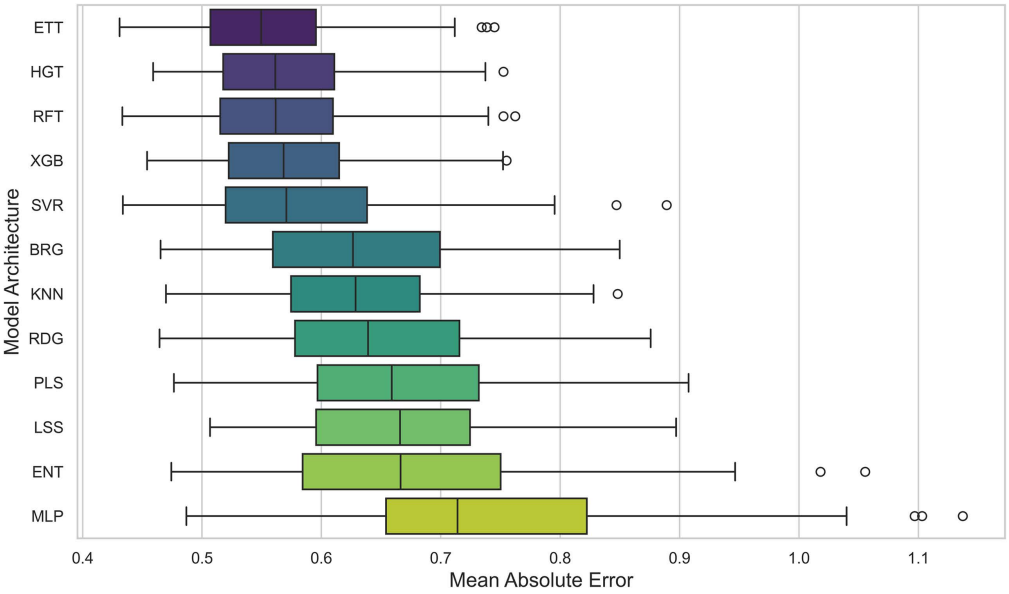

Fig 11. Distribution of MAE for pIC50 Across Model Architectures. Boxplot showing the average MAE of prediction values for pIC50 across multiple cancer therapeutic datasets, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g011

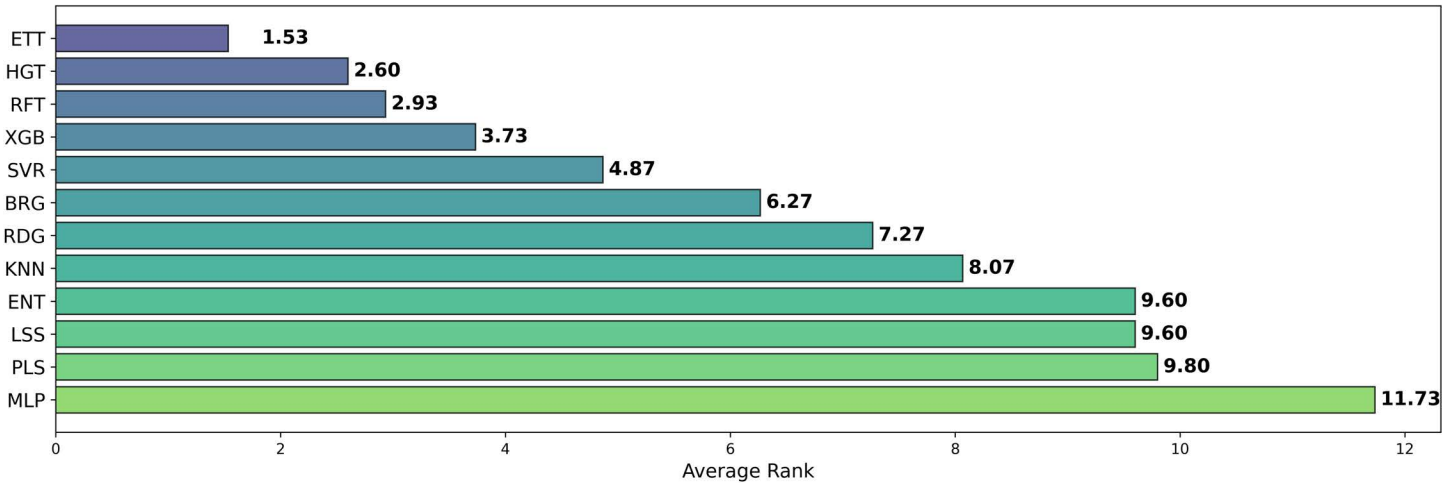

Fig 12. Average Ranking of R^2 for pIC50 Across Model Architectures. Boxplot showing the average ranking based on R^2 for pIC50 across multiple cancer therapeutic datasets and chemical representations, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g012

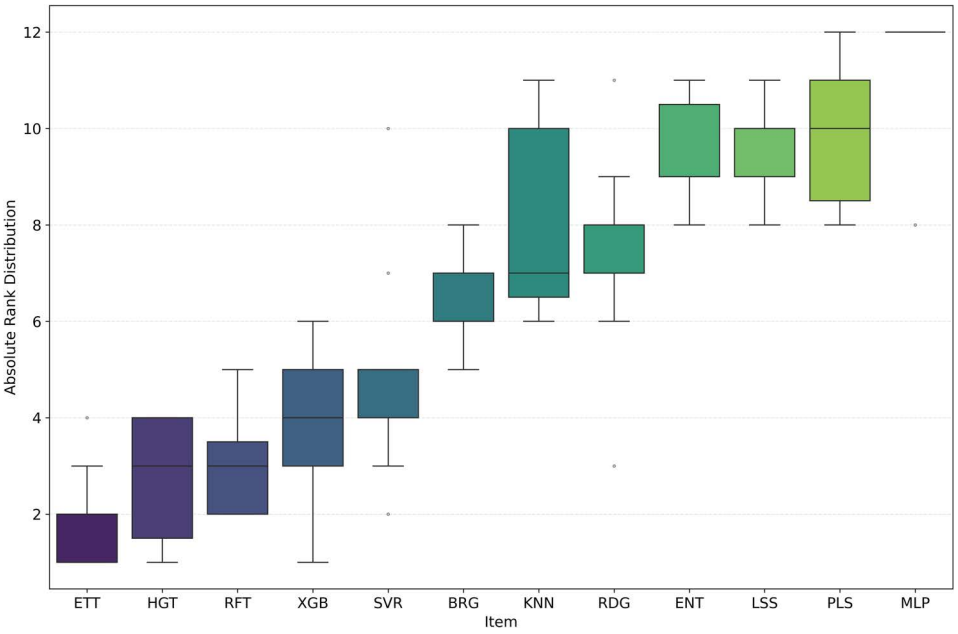

Fig 13. Ranking Distribution of R^2 for pIC50 Across Model Architectures. Boxplot showing the ranking based on R^2 for pIC50 across multiple cancer therapeutic datasets and chemical representations, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g013

The Pearson correlation coefficient between the dataset number and the average R^2 was −0.03, which means no strong correlation was found between the dataset size and the R^2 value. Significant effects of the dataset size on the performance metrics have been reported in classification tasks, as small datasets are not capable of capturing the population features, leading to overfitting, bias, poor generalization capabilities, and, in some cases, inaccurate predictions [84].

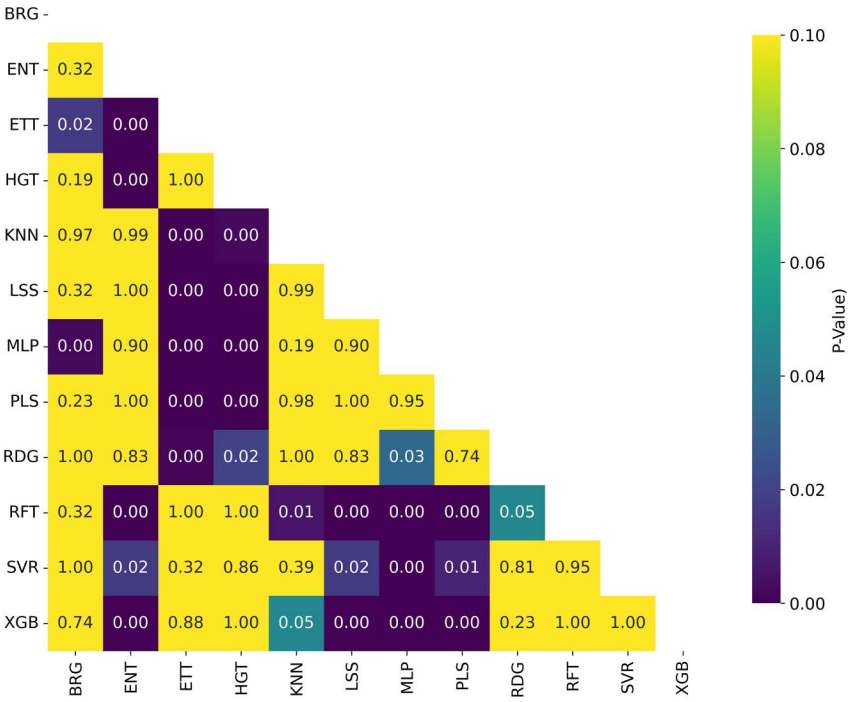

Fig 14. Distribution of Nemenyi p-values for Ranking Statistical Significance Across Model Architectures. Heatmap showing Nemenyi p-values for ranking statistical significance across multiple cancer therapeutic datasets and chemical representations, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g014

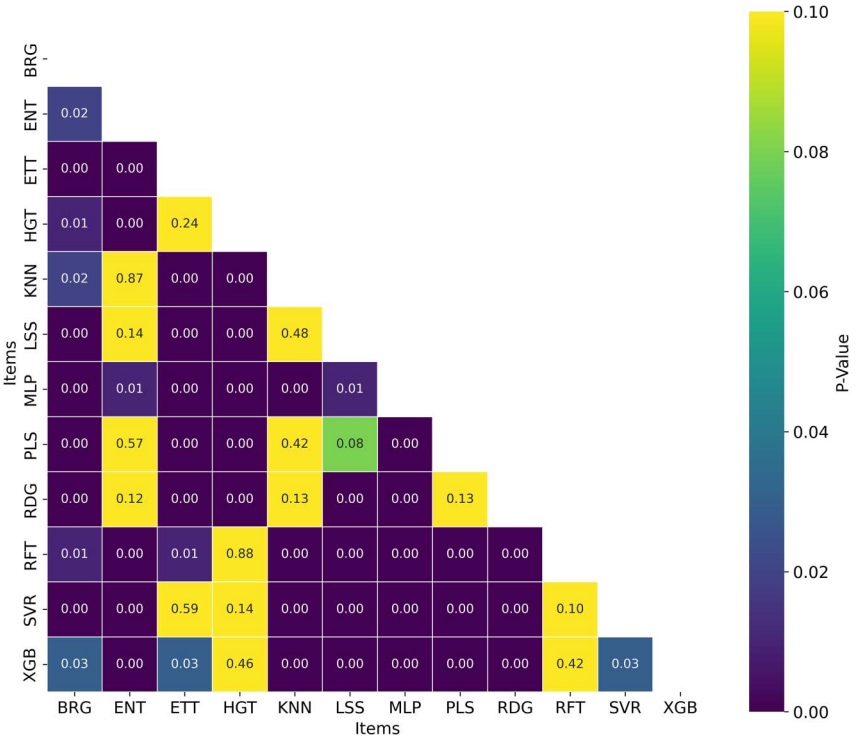

Fig 15. Distribution of T-test p-values for Statistical Significance of R^2 Across Model Architectures. Heatmap showing T-test p-values for statistical significance of R^2 across multiple cancer therapeutic datasets and chemical representations, organized by model architecture.

https://doi.org/10.1371/journal.pone.0343654.g015

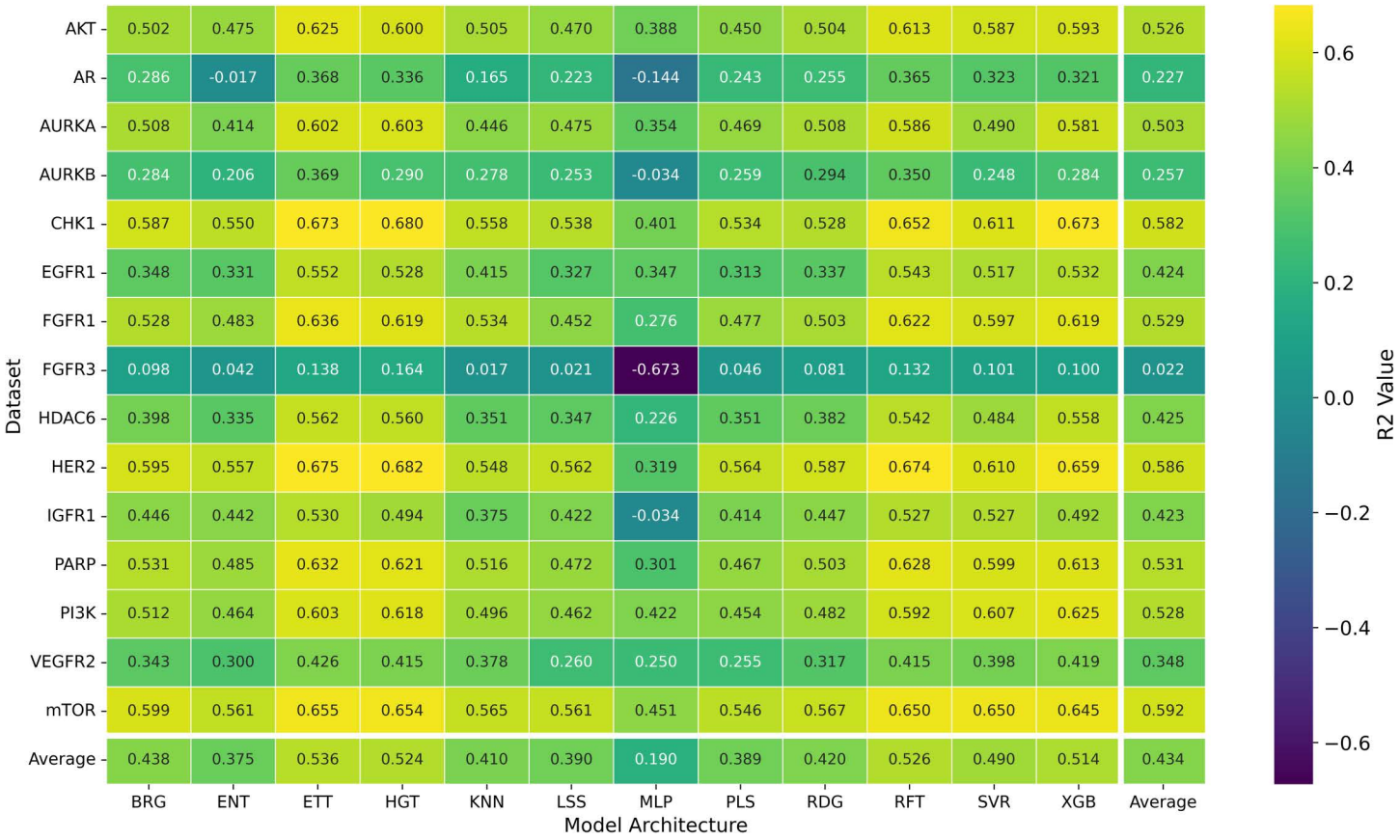

Fig 16. Distribution of R^2 Prediction Values for pIC50 Across Datasets and Model Architectures. Heatmap showing the average R^2 prediction values for pIC50 across multiple cancer therapeutic datasets, organized by chemical representations and model architectures.

https://doi.org/10.1371/journal.pone.0343654.g016

Further dataset analysis could enhance predictive performance, as the quality and content of the database can significantly influence the quality and validity of the QSAR model [85].

Finally, the variability in model performance across different cancer targets, ranging from an R^2 of 0.0.592 for mTOR to 0.227 for AR, underscores the necessity of defining a rigorous Applicability Domain (AD) before deploying these models prospectively. As per OECD principles, a model's predictive reliability is strictly limited to the chemical space defined by the training data [86]. The high disagreement observed in the AR and FGFR3 datasets suggests the presence of structural outliers or diverse binding modes that the models failed to generalize. Future implementation of these QSAR pipelines must incorporate AD assessment techniques, such as distance-to-model metrics or probability density estimation, to filter out unreliable predictions for compounds that are structurally dissimilar to the training set [87].

While this study provides valuable insights into the relationship between molecular representations and machine learning models in predicting bioactivity, several limitations may have influenced the findings. First, the datasets used, while relevant, are limited in size and chemical diversity. With only 15 cancer-related therapeutic targets, the data may only partially capture the broader chemical space, which could limit the generalizability of the models. A more diverse set of biological targets, combined with larger datasets, would be essential to improve the robustness of the results. Additionally, although RDF demonstrated strong performance, other molecular representations lagged. This may stem from their inability to capture molecular details as comprehensively as RDF, AVN and, ECFP6. Exploring other modern representations,

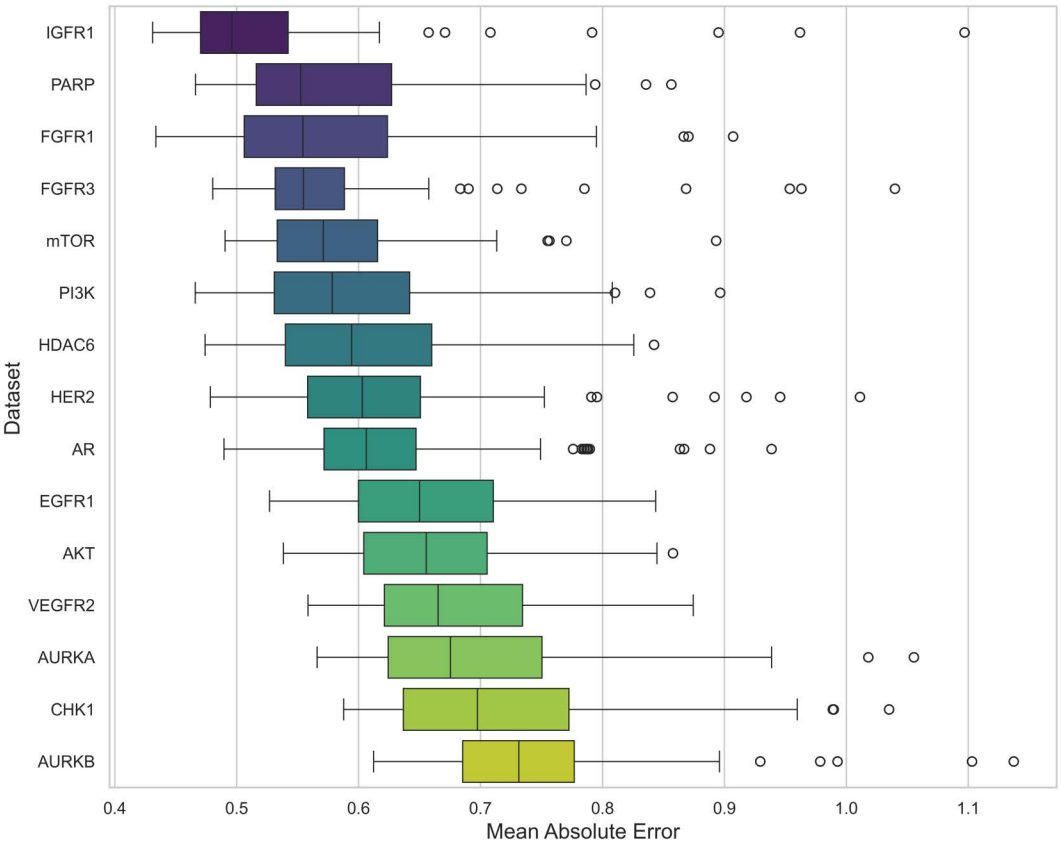

Fig 17. Distribution of MAE for pIC50 Across Datasets. Boxplot showing the average MAE of prediction values for pIC50 across multiple cancer therapeutic datasets, organized by dataset.

https://doi.org/10.1371/journal.pone.0343654.g017

such as 3D descriptors or graph neural network-based representations, could provide a broader evaluation of their impact on model performance.

Furthermore, the study focused primarily on traditional machine learning models, such as Random Forest and Support Vector Regressor, which may not fully exploit the complexity of molecular data compared to more advanced technique s like deep learning models. Future research could benefit from incorporating these newer models, such as graph neural networks or convolutional neural networks, which can capture more nuanced molecular interactions.

We acknowledge the importance of experimental validation in ensuring the reliability of bioactivity predictions. As noted, the current study relies on computational predictions based on available datasets such as ChEMBL, which, while informative, lack experimental confirmation. The absence of comparison with recognized inhibitors is indeed a limitation of the study. To address this concern, future work will include experimental validation, such as molecular docking studies or in vitro testing, to help validate the bioactivity predictions and explore potential mechanisms of action.

Conclusion

After examining the impact of chemical representations in QSAR models, we conclude that RDF exhibited the best performance, followed by AVN and ECFP6. APC showed the least favorable performance, recommending their use as a supplement in combination with other representations. The performance of each representation may be influenced by other

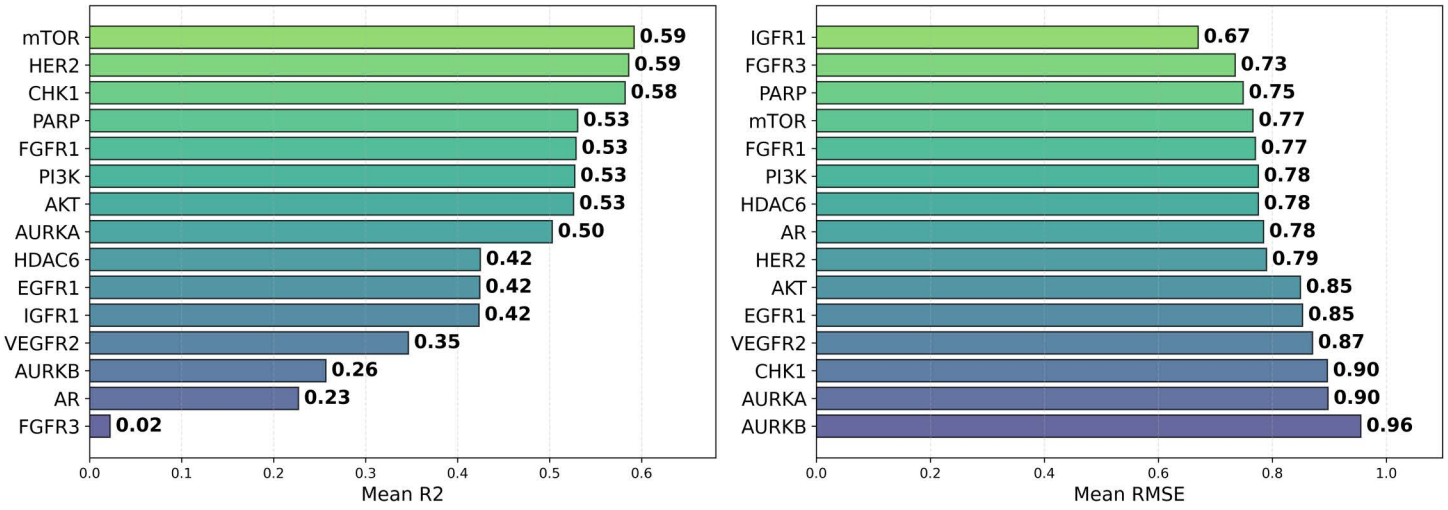

Fig 18. Distribution of R^2 and RMSE for pIC50 Across Datasets. Boxplot (left) showing average R^2 and (right) MAE of prediction values for pIC50 across multiple cancer therapeutic datasets, organized by dataset.

https://doi.org/10.1371/journal.pone.0343654.g018

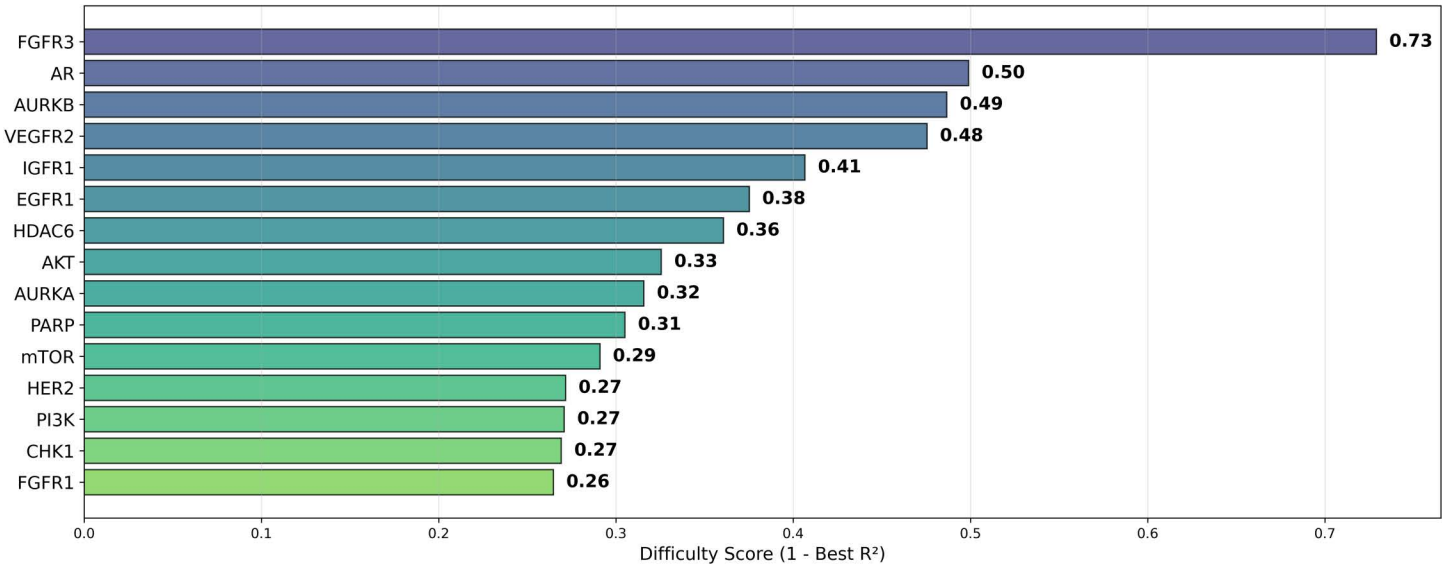

Fig 19. Difficulty Score (1 – R^2) for pIC50 Across Datasets. Boxplot showing the difficulty score (1 – R^2) for pIC50 across model architectures and chemical representations, organized by cancer therapeutic datasets.

https://doi.org/10.1371/journal.pone.0343654.g019

factors such as the choice of machine learning model, the dataset used, and the nature of the task. While RDF appears robust, it is essential to consider alternative representations and combinations, such as using descriptors as supplements to enhance model performance.

Regarding machine learning models, the ETT algorithm demonstrated the highest effectiveness, followed by HGT and RFT, with no statistically significant differences found between ETT and HGT/SVR, indicating that these algorithms might be more reliable for QSAR modeling. The dataset used also affected QSAR model performance, likely related to aspects

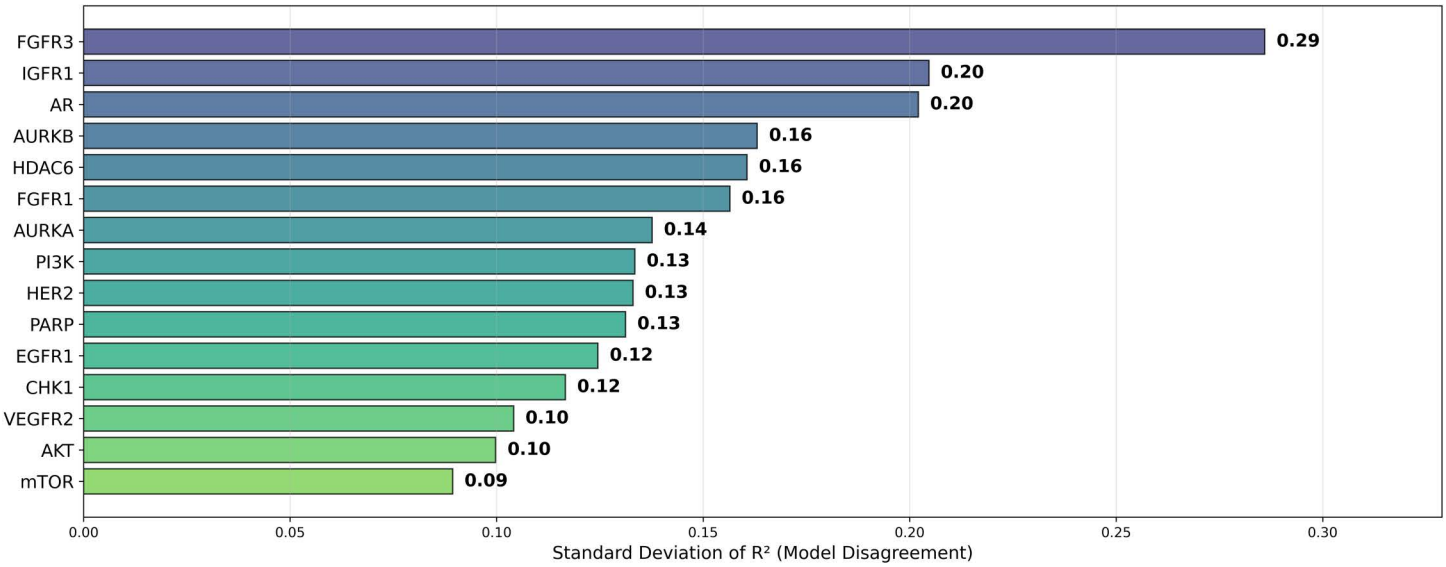

Fig 20. Dataset Disagreement (Standard Deviation) of Average R^2 for pIC50 Across Datasets. Boxplot showing model disagreement (standard deviation) of average R^2 for pIC50 across model architectures and chemical representations, organized by cancer therapeutic datasets.

https://doi.org/10.1371/journal.pone.0343654.g020

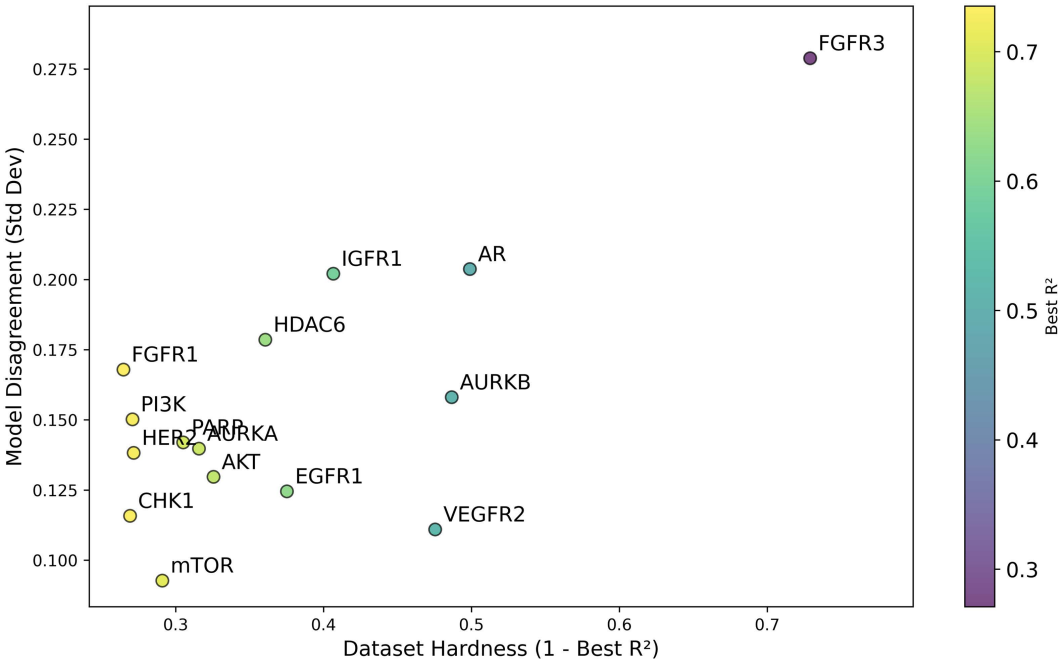

Fig 21. Dataset Disagreement (Standard Deviation) of Average R^2 for pIC50 Across Datasets. Scatterplot showing the relationship between model difficulty (1 – best R^2) and model disagreement (standard deviation) of average R^2 for pIC50 across model architectures and chemical representations, organized by cancer therapeutic datasets.

https://doi.org/10.1371/journal.pone.0343654.g021

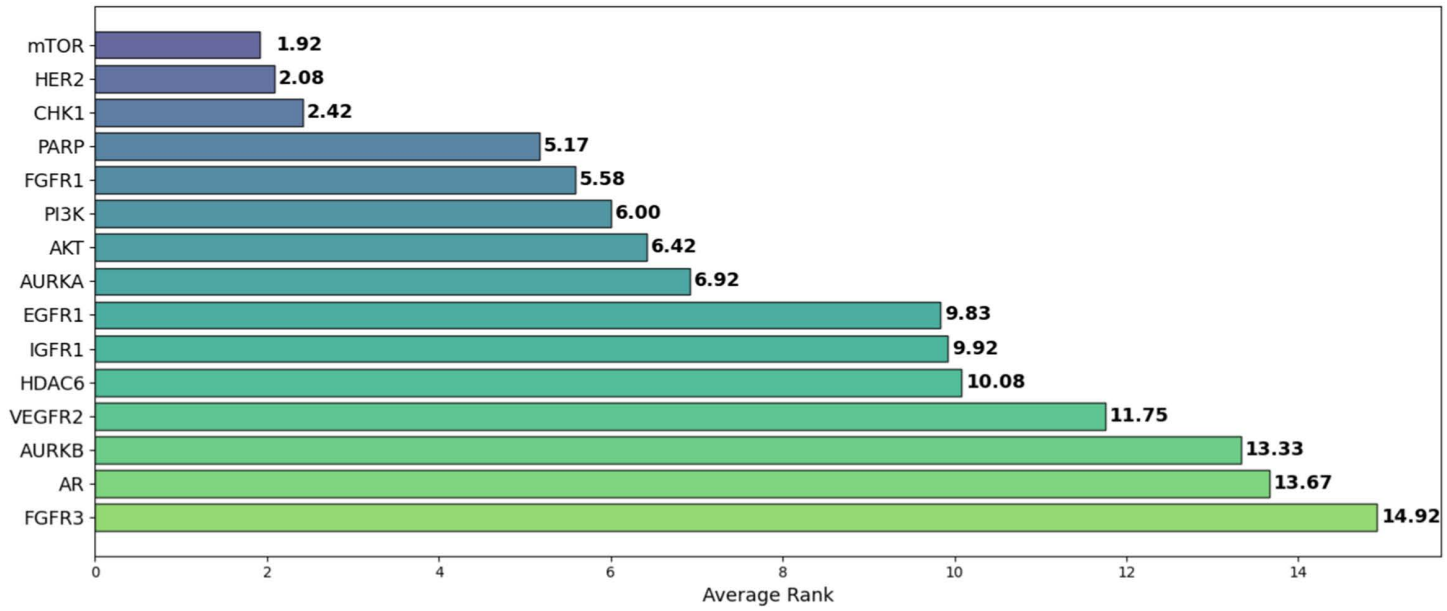

Fig 22. Average Ranking of R^2 for pIC50 Across Datasets. Boxplot showing the average ranking based on R^2 for pIC50 across model architectures and chemical representations, organized by cancer therapeutic datasets.

https://doi.org/10.1371/journal.pone.0343654.g022

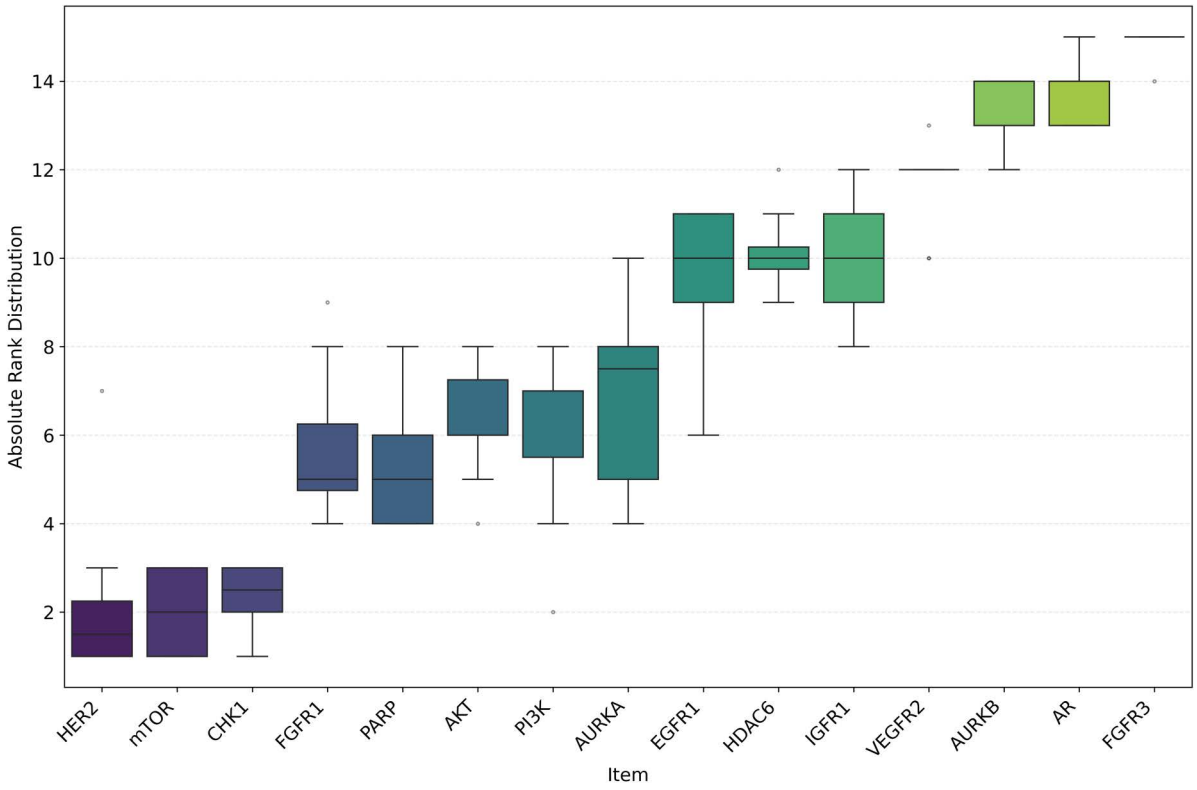

Fig 23. Ranking Distribution of R^2 for pIC50 Across Datasets. Boxplot showing the ranking based on R^2 for pIC50 across model architectures and chemical representations, organized by cancer therapeutic datasets.

https://doi.org/10.1371/journal.pone.0343654.g023

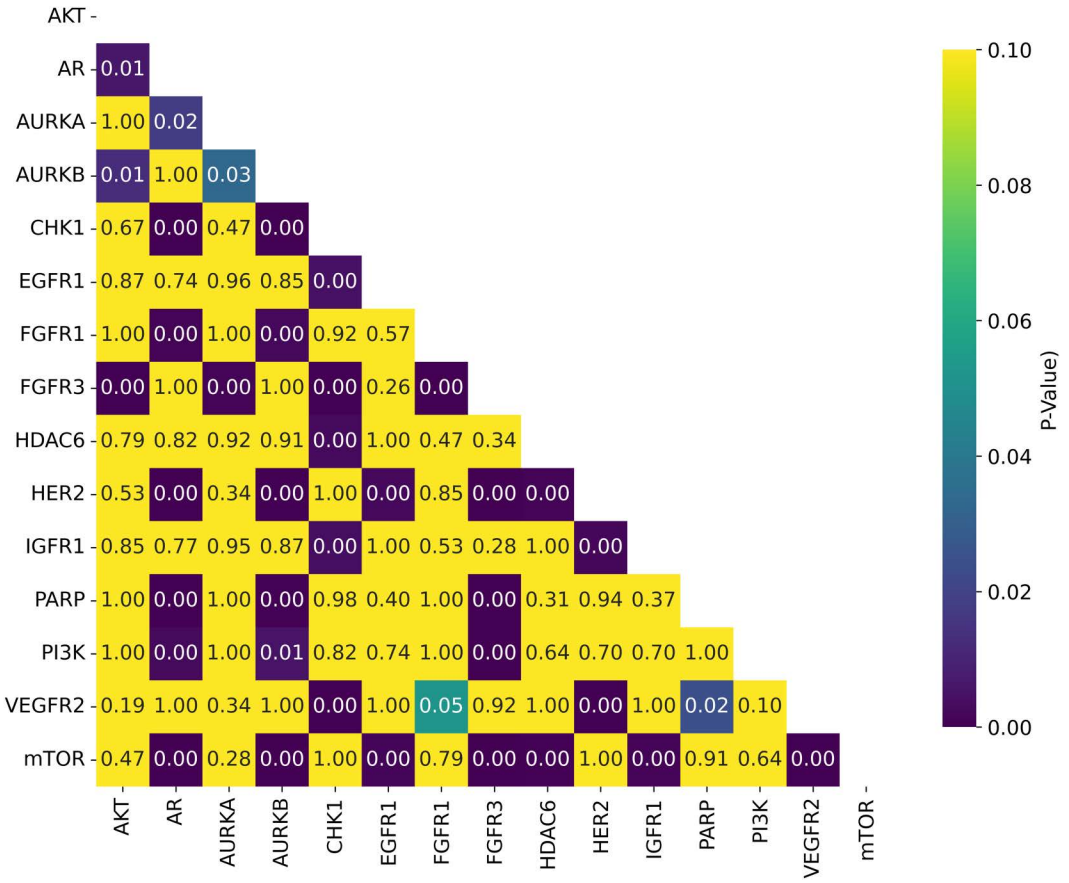

Fig 24. Distribution of Nemenyi p-values for Ranking Statistical Significance Across Datasets. Heatmap showing the Nemenyi p-values for ranking statistical significance across multiple cancer therapeutic datasets, organized by chemical representations and model architectures.

https://doi.org/10.1371/journal.pone.0343654.g024

such as quality and validity. A strong correlation was observed with AVN/ECFP4/ECFP6 chemical representations on ETT/HGT/SVR ($R^2 \geq 0.7$). The acceptance of the alternative hypothesis in many cases suggests that QSAR model conditions (like representation language, preprocessing, and dataset) significantly impact performance, highlighting their importance in research.

Future work in QSAR modeling should focus on exploring novel chemical representations, such as 3D molecular descriptors, and advanced deep learning techniques like graph neural networks (GNNs) or convolutional neural networks (CNNs), which have the potential to capture intricate molecular interactions and spatial properties more effectively. Additionally, integrating multi-omics data (transcriptomics and proteomics) could significantly improve bioactivity predictions' accuracy and contextual relevance. Ensemble modeling approaches could enhance model robustness and generalization, such as stacking different machine learning models and transfer learning techniques, which leverage pre-trained models on related tasks. Tackling issues like data imbalance and bias, incorporating explainable AI techniques to improve model interpretability, and exploring emerging areas such as quantum descriptors and quantum machine learning are promising directions for advancing QSAR research.

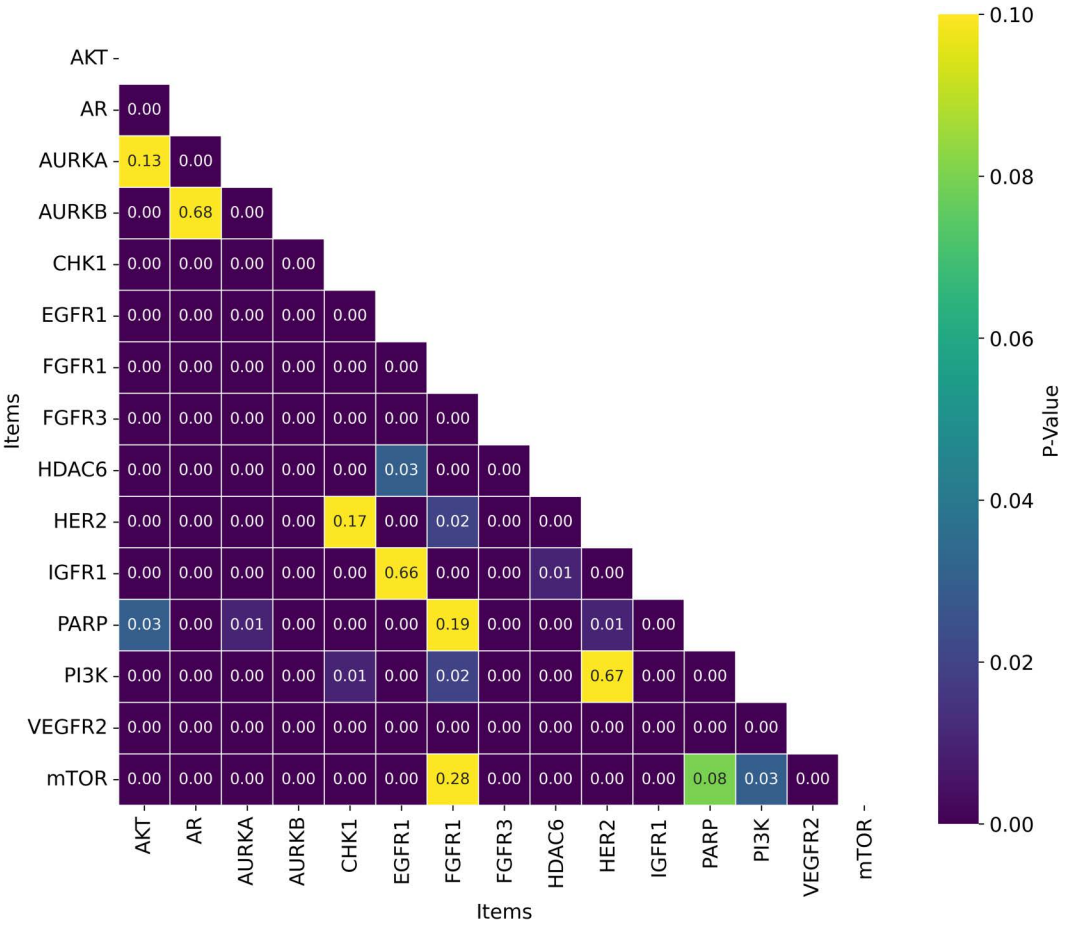

Fig 25. Distribution of T-test p-values for Statistical Significance of R^2 Across Datasets. Heatmap showing the T-test p-values for R^2 statistical significance across multiple cancer therapeutic datasets, organized by chemical representations and model architectures.

https://doi.org/10.1371/journal.pone.0343654.g025

Table 3. Comparison of R^2 and RMSE Performance Across Different Models. Values represent the coefficient of determination (R^2) and RMSE. The highest-performing model (highest R^2 and lowest RMSE) is indicated in bold.

Model	R^2	RMSE
BRG	0.4424	0.8027
ENT	0.3807	0.8538
ETT	**0.5408**	**0.7275**
HGT	0.5301	0.7408
*K*NN	0.4153	0.8348
LSS	0.3942	0.8458
MLP	0.2044	0.9668
PLS	0.3935	0.8489
RDG	0.4236	0.8282
RFT	0.5305	0.7346
SVR	0.4973	0.7534
XGB	0.5212	0.7447

https://doi.org/10.1371/journal.pone.0343654.t003

Table 4. Comparison of R^2 and RMSE Performance Across Different Datasets. Values represent the coefficient of determination (R^2) and RMSE. The highest-performing dataset (highest R^2 and lowest RMSE) is indicated in bold.

Dataset	R^2	RMSE
HER2	0.5861	0.7898
HDAC6	0.4247	0.7756
AR	0.2270	0.7845
IGFR1	0.4233	**0.6699**
EGFR1	0.4241	0.8529
AURKB	0.2568	0.9552
FGFR3	0.0223	0.7347
VEGFR2	0.3466	0.8704
mTOR	**0.5920**	0.7661
PARP	0.5305	0.7490
PI3K	0.5273	0.7755
FGFR1	0.5288	0.7703
PI3K	0.5273	0.7755
AKT	0.5261	0.8493
CHK1	0.5822	0.8966
AURKA	0.5029	0.8975

https://doi.org/10.1371/journal.pone.0343654.t004

Acknowledgments

We extend our sincere gratitude to Tecnologico de Monterrey for their commitment to open access publishing, demonstrated by their support in covering the Article Processing Charge (APC) of this work.

Author contributions

Conceptualization: Patricio Adrian Zapata.

Data curation: Raúl Acosta Murillo.

Formal analysis: Raúl Acosta Murillo.

Investigation: José Carlos Ortiz-Bayliss, Patricio Adrian Zapata.

Methodology: Raúl Acosta Murillo, José Carlos Ortiz-Bayliss, Patricio Adrian Zapata.

Supervision: José Carlos Ortiz-Bayliss, Patricio Adrian Zapata.

Writing – original draft: Raúl Acosta Murillo.

Writing – review & editing: José Carlos Ortiz-Bayliss, Patricio Adrian Zapata.

References

1. Roy K, Kar S, Das RN. QSAR/QSPR modeling: introduction. Springer Briefs in Molecular Science. 2015. p. 1–36.
2. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. WIREs Comput Mol Sci. 2022;12(5). https://doi.org/10.1002/wcms.1603
3. Koirala M, Yan L, Mohamed Z, DiPaola M. AI-integrated QSAR modeling for enhanced drug discovery: from classical approaches to deep learning and structural insight. Int J Mol Sci. 2025;26(19):9384. https://doi.org/10.3390/ijms26199384 PMID: 41096653
4. Niazi SK, Mariam Z. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. Int J Mol Sci. 2023;24(14):11488. https://doi.org/10.3390/ijms241411488 PMID: 37511247

5. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525–64. https://doi.org/10.1039/d0cs00098a PMID: 32356548
6. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023;73(1):17–48. https://doi.org/10.3322/caac.21763 PMID: 36633525
7. Instituto Mexicano del Seguro Social. Epidemiología del cáncer de mama. 2022. https://www.gob.mx/imss/articulos/epidemiologia-del-cancer-de-mama-318014
8. Ye F, Dewanjee S, Li Y, Jha NK, Chen Z-S, Kumar A, et al. Advancements in clinical aspects of targeted therapy and immunotherapy in breast cancer. Mol Cancer. 2023;22(1):105. https://doi.org/10.1186/s12943-023-01805-y PMID: 37415164
9. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42(6):1273–80. https://doi.org/10.1021/ci010132r PMID: 12444722
10. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci. 1985;25(2):64–73. https://doi.org/10.1021/ci00046a002
11. Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. J Chem Inf Comput Sci. 1987;27(2):82–5. https://doi.org/10.1021/ci00054a008
12. Gedeck P, Rohde B, Bartels C. QSPR with the Avalon Fingerprint Module of the KNIME Chemistry Extensions. J Chem Inf Model. 2006;46(5):1924–36. https://doi.org/10.1021/ci060135t
13. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54. https://doi.org/10.1021/ci100050t PMID: 20426451
14. Wang Z, Chen J, Hong H. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. Environ Sci Technol. 2021;55(10):6857–66. https://doi.org/10.1021/acs.est.0c07040 PMID: 33914508
15. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005 PMID: 25132639
16. Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform. 2020;12(1):43. https://doi.org/10.1186/s13321-020-00445-4 PMID: 33431010
17. Rensi SE, Altman RB. Shallow Representation Learning via Kernel PCA Improves QSAR Modelability. J Chem Inf Model. 2017;57(8):1859–67. https://doi.org/10.1021/acs.jcim.6b00694 PMID: 28727421
18. Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model. 2018;58(1):27–35. https://doi.org/10.1021/acs.jcim.7b00616 PMID: 29268609
19. Sabando MV, Ponzoni I, Milios EE, Soto AJ. Using molecular embeddings in QSAR modeling: does it make a difference?. Brief Bioinform. 2022;23(1):bbab365. https://doi.org/10.1093/bib/bbab365 PMID: 34498670
20. Daoui O, Elkhattabi S, Chtita S, Elkhalabi R, Zgou H, Benjelloun AT. QSAR, molecular docking and ADMET properties in silico studies of novel 4,5,6,7-tetrahydrobenzo[D]-thiazol-2-Yl derivatives derived from dimedone as potent anti-tumor agents through inhibition of C-Met receptor tyrosine kinase. Heliyon. 2021;7(7):e07463. https://doi.org/10.1016/j.heliyon.2021.e07463 PMID: 34296007
21. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4. https://doi.org/10.1186/s13321-018-0258-y PMID: 29411163
22. Riniker S, Landrum GA. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. J Chem Inf Model. 2015;55(12):2562–74. https://doi.org/10.1021/acs.jcim.5b00654 PMID: 26575315
23. Kuhn M, Johnson K. Applied Predictive Modeling. Springer New York. 2013. https://doi.org/10.1007/978-1-4614-6849-3
24. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010;29(6–7):476–88. https://doi.org/10.1002/minf.201000061 PMID: 27463326
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.
26. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634
27. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
28. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. https://doi.org/10.1023/a:1010933404324
29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 785–94.
30. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/bf00994018
31. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. The MIT Press; 2005. http://dx.doi.org/10.7551/mitpress/3206.001.0001
32. Chakravarti SK, Alla SRM. Descriptor Free QSAR Modeling Using Deep Learning With Long Short-Term Memory Neural Networks. Front Artif Intell. 2019;2:17. https://doi.org/10.3389/frai.2019.00017 PMID: 33733106

33. Jiang D, Wu Z, Hsieh C-Y, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12. https://doi.org/10.1186/s13321-020-00479-8 PMID: 33597034
34. Broccatelli F, Trager R, Reutlinger M, Karypis G, Li M. Benchmarking Accuracy and Generalizability of Four Graph Neural Networks Using Large In Vitro ADME Datasets from Different Chemical Spaces. Mol Inform. 2022;41(8):e2100321. https://doi.org/10.1002/minf.202100321 PMID: 35156325
35. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. 2020. https://arxiv.org/abs/2010.09885
36. Méndez-Lucio O, Nicolaou CA, Earnshaw B. MolE: a foundation model for molecular graphs using disentangled attention. Nat Commun. 2024;15(1):9431. https://doi.org/10.1038/s41467-024-53751-y PMID: 39532853
37. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 2017. https://arxiv.org/abs/1705.07874
38. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D1180–92. https://doi.org/10.1093/nar/gkad1004 PMID: 37933841
39. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010;50(7):1189–204. https://doi.org/10.1021/ci100176x PMID: 20572635
40. Landrum G, Tosco P, Kelley B, Ric G, Cosgrove D, Sriniker S. rdkit/rdkit: 2023_09_4 (Q3 2023) release. 2024.
41. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2017;9(2):513–30. https://doi.org/10.1039/c7sc02664a PMID: 29629118
42. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform. 2013;5(1):26. https://doi.org/10.1186/1758-2946-5-26 PMID: 23721588
43. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer New York. 2009. https://doi.org/10.1007/978-0-387-84858-7
44. Habash M, Alshakhshir S, Awwad S, Abu-Samak M. The discovery of potential tumor necrosis factor alpha converting enzyme inhibitors via implementation of K Nearest Neighbor QSAR analysis. PHAR. 2023;70(2):247–61. https://doi.org/10.3897/pharmacia.70.e96423
45. Banerjee A, Roy K. Machine-learning-based similarity meets traditional QSAR: "q-RASAR" for the enhancement of the external predictivity and detection of prediction confidence outliers in an hERG toxicity dataset. Chemometrics and Intelligent Laboratory Systems. 2023;237:104829. https://doi.org/10.1016/j.chemolab.2023.104829
46. Abbod M, Mohammad A. Combined interaction of fungicides binary mixtures: experimental study and machine learning-driven QSAR modeling. Sci Rep. 2024;14(1):12700. https://doi.org/10.1038/s41598-024-63708-2 PMID: 38830957
47. Bahia MS, Kaspi O, Touitou M, Binayev I, Dhail S, Spiegel J, et al. A comparison between 2D and 3D descriptors in QSAR modeling based on bio-active conformations. Mol Inform. 2023;42(4):e2200186. https://doi.org/10.1002/minf.202200186 PMID: 36617991
48. Takeda K, Takeuchi K, Sakuratani Y, Kimbara K. Optimal selection of learning data for highly accurate QSAR prediction of chemical biodegradability: a machine learning-based approach. SAR QSAR Environ Res. 2023;34(9):729–43. https://doi.org/10.1080/1062936X.2023.2251889 PMID: 37674414
49. Kramer O. K-Nearest Neighbors. Intelligent Systems Reference Library. Springer Berlin Heidelberg. 2013. p. 13–23. https://doi.org/10.1007/978-3-642-38652-7_2
50. Mehmood T, Sæbø S, Liland KH. Comparison of variable selection methods in partial least squares regression. Journal of Chemometrics. 2020;34(6). https://doi.org/10.1002/cem.3226
51. Awad M, Khanna R. Support Vector Regression. Efficient Learning Machines. Apress. 2015. p. 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4
52. Rigatti SJ. Random Forest. J Insur Med. 2017;47(1):31–9. https://doi.org/10.17849/insm-47-01-31-39.1 PMID: 28836909
53. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x
54. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1
55. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017;30:3146–54.
56. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6. https://doi.org/10.1038/323533a0
57. MacKay DJC. Bayesian Interpolation. Neural Computation. 1992;4(3):415–47. https://doi.org/10.1162/neco.1992.4.3.415
58. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 1967;13(1):21–7. https://doi.org/10.1109/tit.1967.1053964
59. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. J Mach Learn Res. 2017;18(1):6765–816.
60. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem. 1996;39(15):2887–93. https://doi.org/10.1021/jm9602928 PMID: 8709122

61. Kumari M, Alfagham AT, Elgorban AM, Mallik S, Lemos B, Ray K. Development of convolutional neural network-based QSAR model to predict cardiotoxicity and principal component analysis, fingerprint analysis. J King Saud Univ Sci. 2025;37:3112024. https://doi.org/10.25259/jksus_311_2024
62. Abdallah RM, Hasan HE, Hammad A. Predictive modeling of skin permeability for molecules: Investigating FDA-approved drug permeability with various AI algorithms. PLOS Digit Health. 2024;3(4):e0000483. https://doi.org/10.1371/journal.pdig.0000483 PMID: 38568888
63. Nguyen T, Bavarian M. A Machine Learning Framework for Predicting the Glass Transition Temperature of Homopolymers. Ind Eng Chem Res. 2022;61(34):12690–8. https://doi.org/10.1021/acs.iecr.2c01302
64. Tukey JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley. 1977.
65. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
66. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543
67. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates. 1988.
68. Shtivelman E, Beer TM, Evans CP. Molecular pathways and targets in prostate cancer. Oncotarget. 2014;5(17):7217–59. https://doi.org/10.18632/oncotarget.2406 PMID: 25277175
69. Orosz Á, Héberger K, Rácz A. Comparison of Descriptor- and Fingerprint Sets in Machine Learning Models for ADME-Tox Targets. Front Chem. 2022;10:852893. https://doi.org/10.3389/fchem.2022.852893 PMID: 35755260
70. Lee M, Min K. A Comparative Study of the Performance for Predicting Biodegradability Classification: The Quantitative Structure-Activity Relationship Model vs the Graph Convolutional Network. ACS Omega. 2022;7(4):3649–55. https://doi.org/10.1021/acsomega.1c06274 PMID: 35128273
71. Xie L, Xu L, Kong R, Chang S, Xu X. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. Front Pharmacol. 2020;11:606668. https://doi.org/10.3389/fphar.2020.606668 PMID: 33488387
72. Zagidullin B, Wang Z, Guan Y, Pitkänen E, Tang J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. Brief Bioinform. 2021;22(6):bbab291. https://doi.org/10.1093/bib/bbab291 PMID: 34401895
73. Hemmer MC, Steinhauer V, Gasteiger J. Deriving the 3D structure of organic molecules from their infrared spectra. Vibrational Spectroscopy. 1999;19(1):151–64. https://doi.org/10.1016/s0924-2031(99)00014-4
74. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH. 2000.
75. De Vivo M, Bottegoni G, Berteotti A, Recanatini M, Gervasio FL, Cavalli A. Cyclin-dependent kinases: bridging their structure and function through computations. Future Med Chem. 2011;3(12):1551–9. https://doi.org/10.4155/fmc.11.113 PMID: 21882947
76. Winter R, Montanari F, Noé F, Clevert D-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci. 2018;10(6):1692–701. https://doi.org/10.1039/c8sc04175j PMID: 30842833
77. Franz F, Daday C, Gräter F. Advances in molecular simulations of protein mechanical properties and function. Curr Opin Struct Biol. 2020;61:132–8. https://doi.org/10.1016/j.sbi.2019.12.015 PMID: 31954324
78. Du Z, Wang D, Li Y. Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. ACS Omega. 2022;7(29):25760–71. https://doi.org/10.1021/acsomega.2c03062 PMID: 35910147
79. Wiriyarattanakul A, Xie W, Toopradab B, Wiriyarattanakul S, Shi L, Rungrotmongkol T, et al. Comparative Study of Machine Learning-Based QSAR Modeling of Anti-inflammatory Compounds from Durian Extraction. ACS Omega. 2024;9(7):7817–26. https://doi.org/10.1021/acsomega.3c07386 PMID: 38405441
80. Tsou LK, Yeh S-H, Ueng S-H, Chang C-P, Song J-S, Wu M-H, et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. Sci Rep. 2020;10(1):16771. https://doi.org/10.1038/s41598-020-73681-1 PMID: 33033310
81. Diéguez-Santana K, Casañola-Martin GM, Torres R, Rasulev B, Green JR, González-Díaz H. Machine Learning Study of Metabolic Networks vs ChEMBL Data of Antibacterial Compounds. Mol Pharm. 2022;19(7):2151–63. https://doi.org/10.1021/acs.molpharmaceut.2c00029 PMID: 35671399
82. Lane TR, Foil DH, Minerali E, Urbina F, Zorn KM, Ekins S. Bioactivity Comparison across Multiple Machine Learning Algorithms Using over 5000 Datasets for Drug Discovery. Mol Pharm. 2021;18(1):403–15. https://doi.org/10.1021/acs.molpharmaceut.0c01013 PMID: 33325717
83. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Adv Neural Inf Process Syst. 2022;35:507–20.
84. Sutojo T, Rustad S, Akrom M, Syukur A, Shidik GF, Dipojono HK. A machine learning approach for corrosion small datasets. npj Mater Degrad. 2023;7(1). https://doi.org/10.1038/s41529-023-00336-7
85. Burgoon LD, Kluxen FM, Hüser A, Frericks M. The database makes the poison: How the selection of datasets in QSAR models impacts toxicant prediction of higher tier endpoints. Regul Toxicol Pharmacol. 2024;151:105663. https://doi.org/10.1016/j.yrtph.2024.105663 PMID: 38871173
86. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim. 2005;33(2):155–73. https://doi.org/10.1177/026119290503300209 PMID: 16180989
87. Kar S, Roy K. Applicability domain: a step toward confident predictions and acceptability of QSAR models. Molecular Diversity. 2018;22:763–82. https://doi.org/10.1007/s11030-018-9878-0