Benchmarking of small molecule feature representations for hERG, Nav1.5, and Cav1.2 cardiotoxicity prediction

Benchmarking of Small Molecule Feature Representations for hERG, Nav1.5, and Cav1.2 Cardiotoxicity Prediction

Issar Arab [1,2 *], Kristof Egghe[1], Kris Laukens [1,2], Ke Chen [3], Khaled Barakat [4], Wout Bittremieux [1,2]

[1]Department of Computer Science, University of Antwerp, 2020 Antwerp, Belgium

[2]Biomedical Informatics Network Antwerpen (biomina), 2020 Antwerp, Belgium

[3]Chair for Theoretical Chemistry, Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany

[4]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta 8613, Canada

Abstract

In the field of drug discovery, there is a substantial challenge in seeking out chemical structures that possess desirable pharmacological, toxicological, and pharmacokinetic properties. Complications arise when drugs interfere with the functioning of cardiac ion channels, leading to serious cardiovascular consequences. The discontinuation and removal of numerous approved drugs from the market or at late development stages in the pipeline due to such inhibitory effects further highlight the urgency of addressing this issue. Consequently, the early prediction of potential blockers targeting cardiac ion channels during the drug discovery process is of paramount importance. This study introduces a deep learning framework that computationally determines the cardiotoxicity associated with the voltage-gated potassium channel (hERG), the voltage-gated calcium channel (Cav1.2), and the voltage-gated sodium channel (Nav1.5) for drug candidates. The predictive capabilities of three feature representations—molecular fingerprints, descriptors, and graph-based numerical representations—are rigorously benchmarked. Additionally, a novel training and evaluation dataset framework is presented, enabling predictive model training of drug off-target cardiotoxicity using a comprehensive and large curated dataset covering these three cardiac ion channels. To facilitate these predictions, a robust and comprehensive small molecule cardiotoxicity prediction tool named CToxPred has been developed. It is made available as open source under the permissive MIT license at https://github.com/issararab/CToxPred.

1. Introduction

Drug discovery is a complex and multifaceted process that involves the identification and development of new therapeutic agents to treat various diseases. It encompasses a range of scientific disciplines, including chemistry, biology, pharmacology, and medicine, and it involves the discovery, design, synthesis, and optimization of new chemical entities that can modulate disease targets in a safe and effective manner. Conventionally, within the realm of drug discovery, researchers undertake *in vitro* and *in vivo* studies to assess the pharmacodynamics and pharmacokinetic (PD/PK) characteristics of chosen candidates derived from preliminary screening outcomes [1][2]. These experiments are not only demanding in terms of time and financial resources but also raise ethical concerns, particularly in cases involving animal testing [3]. Previous research has demonstrated that the creation of new drugs is a time-consuming and elaborate process that on average can take from six to twelve years and involve expenses of up to 2.5 billion dollars [4][5][6]. Out of this substantial

financial cost, approximately 1.1 billion dollars is allocated for the stages of drug development preceding human trials [6].

Among the five essential pharmacokinetic attributes—chemical absorption, distribution, metabolism, excretion, and toxicity—is toxicity, which necessitates strict validation prior to granting clinical trial approval for a novel drug candidate [7]. According to the directives outlined by the International Conference on Harmonization of Technical Requirements for the Registration of Pharmaceuticals for Human Use, there should be a preclinical assessment of cardiac ion channels inhibition and QT interval (i.e. time duration between the start of the Q wave and the end of the T wave on an electrocardiogram) prolongation caused by small compounds[8]. This is defined as cardiotoxicity and involves the inhibition of any of the three cardiac ion channels: the voltage-gated potassium channel (hERG), the voltage-gated calcium channel (Cav1.2), and the voltage-gated sodium channel (Nav1.5). The pharmaceutical industry suffers significant losses due to cardiotoxicities that emerge during the early, preclinical, or clinical stages of drug development, leading to the withdrawal of several drugs from the market and the halting of many drug discovery programs in their pipelines [9]. Examples of such drugs are astemizole, terfenadine, sertindole, grepafloxacin, vardenafil, cisapride, and ziprasidone, which have been withdrawn or severely restricted on the use for the undesirable cardiac toxicity side effects [10][11][12].

An effective solution that offers a prominent alternative to reduce costs and advance the development of lead candidates is the field of computer-aided drug discovery (CADD) [6][13][14]. Toxicity prediction algorithms have recently become a very important component of modern CADD [13], with cardiotoxicity prediction algorithms as the most dominant of these methods.

In recent years, researchers have increasingly utilized machine learning (ML) algorithms to construct and deploy robust models for predicting cardiac ion channels' inhibition. Previous reviews [10] [15] have indicated that during the early 2000s, statistical approaches such as Naïve Bayes, Gaussian processes, expectation–maximization, and partial least squares (PLS) were commonly employed ML algorithms [16][17][18][19][20]. The majority of those published prediction methods used less than a thousand compounds to train their models. Later, there has been a notable shift toward the utilization of random forest (RF), support vector machine (SVM), and deep neural network (DNN) methods [21][22][23][24], using up to 15,000 compounds for training the models. This transition is primarily due to their superior empirical performance in the field, as evidenced by multiple research publications outlined recently [25].

However, there is still important room for improvement of cardiotoxicity prediction. First, almost all published methods during the last 20 years focused only on hERG liability prediction, as there were very few bio-activity data available on the other two cardiac ion channels. Second, while previous research has used various representations such as fingerprints, descriptors, and graph structures separately or in combination, none has conducted a comprehensive benchmark to evaluate their predictive performance on any of the three targets. Third, different models were trained on different datasets lacking a common and unique dataset to use for benchmarking. Even though most papers used similar data sources, different curation techniques and decisions were applied. Fourth, published and evaluated models by some authors cannot be validated for better generalization as many published models show overlap between training and test sets; hence presenting over-optimistic performance results. Illustrative examples include the work of Konda et al. [26], wherein their test set exhibits an approximate 50% overlap in molecular compounds with the training set. Another noteworthy scenario is presented by Liu et al. [27], who introduced a model trained and assessed on the dataset published by Doddareddy et al. [28]. Liu et al.'s findings indicated that although their best model demonstrated enhanced performance on the provided test set (achieving an accuracy of 88%

for [28] versus 91% for [27]), its performance declined when evaluated on external sets, reaching 47% and 58% respectively. Our own evaluation further revealed that more than a quarter of the test data exhibited an overlap with the training set, exceeding 80% in terms of structural similarity.

In this study, we present a framework to perform further analyses on a very large open-access and comprehensive hERG, Nav1.5, and Cav1.2 cardiotoxicity integrated database of small molecules and their activities. We also present a deep learning model used to benchmark different feature sets, namely descriptors, fingerprints, and graph-based representations, for cardiotoxicity prediction. Finally, we benchmark our best models on strictly unique extracted external evaluation sets and provide a robust predictive model for each target.

2. Methods

2.1. Compilation of a Cardiotoxicity Database

The compounds demonstrating inhibitory activity used in this study were sourced from various public data repositories, including the ChEMBL bioactivity database [29][30][31], PubChem [32], BindingDB [33][34], hERGCentral [35], and US patent and literature-derived data [36][37][38][39]. The collected data was split into two primary classes based on the available information: IC50-type and inhibition-type values. The IC50-type measurements encompassed inhibitory activity values expressed as half maximum inhibitory concentration (IC50), half maximum effective concentration (EC50), median effective dose (ED50), inhibitory constant (Ki), or dissociation constant (Kd). Conversely, the inhibition-type measurements comprised percentage inhibition values at specific concentrations. It is crucial to recognize that there are several potential sources of bias and inconsistencies that could impact the ultimate curated database, and hence, the outcomes of the developed models. These sources serve as a constraining factor for any models developed around cardiac ion channels [40]. This encompasses disparities in the experimental assays employed for data collection[40][41]. For instance, Guo et al. [42] demonstrated that limitations in the same patch clamp instrument can yield varying measurements for the same compound and to attain stable drug concentrations, repeated compound additions proved to be critical. Moreover, discrepancies can arise when using standard patch-clamp techniques compared to automated instruments, resulting in threefold shifts in IC50 values for highly hydrophobic compounds[42]. Furthermore, comprehending the distinctions in the final data readout is imperative. Take, for example, the concepts of EC50 and IC50, which, although similar, are not entirely identical. EC50 measures the drug concentration needed to achieve half of the maximal effective response, while IC50 signifies the inhibitor concentration at which a 50% reduction in activity is attained. Additionally, ED50 denotes the dose of a medication that generates a specific effect in 50% of the population receiving that dose[43]. These distinctions become particularly relevant in a clinical context, where derivatives of the tested compound or other ion channels can influence the outcomes of the ion channel under consideration[44]. Taken together, despite rigorous data curation efforts, awareness of potential biases and variations in data sources and readouts is essential when constructing models related to the hERG, Nav1.5, or Cav1.2 channels. These considerations are critical for accurate and reliable research and applications in this field. To reduce such biases and discrepancies, we followed the recent data curation methodology outlined by Sato et al [45]. That is, in the case of inhibition-type entries, a manual examination of assay descriptions was conducted to extract the relevant activity values. Values reported with thresholds other than 50% inhibition (e.g., IC70, IC30, IC20, etc.), raw measurements of current, ratios indicating prolonged QT intervals, and similar values were excluded. Additionally, for hERG data, cross-referencing with the hERGCentral retrieved data was performed, and any erroneous activity values were rectified.

Following this step, the chemical structures within each dataset underwent standardization using the Python packages RDKit (http://www.rdkit.org) and MolVS (https://github.com/mcs07/MolVS). The standardization procedure encompassed selecting the largest fragment, eliminating explicit hydrogens, ionization, and calculating stereochemistry. Next, the compounds were encoded as SMILES (Simplified Molecular Input Line Entry System) strings. Because multiple SMILES strings can correspond to the same structure, and there is no universal approach for generating a canonical SMILES string, we converted the SMILES strings to InChI (International Chemical Identifier) keys [46] using RDKit to identify duplicate compounds. All potency values were transformed into nanomolar (nM) units. For duplicate compounds, the mean value was subsequently computed, while only retaining values falling within the 95% quantile range to exclude excessive outliers. The overall methodology employed to construct the extensive small molecule cardiotoxicity database is illustrated in Figure 1.

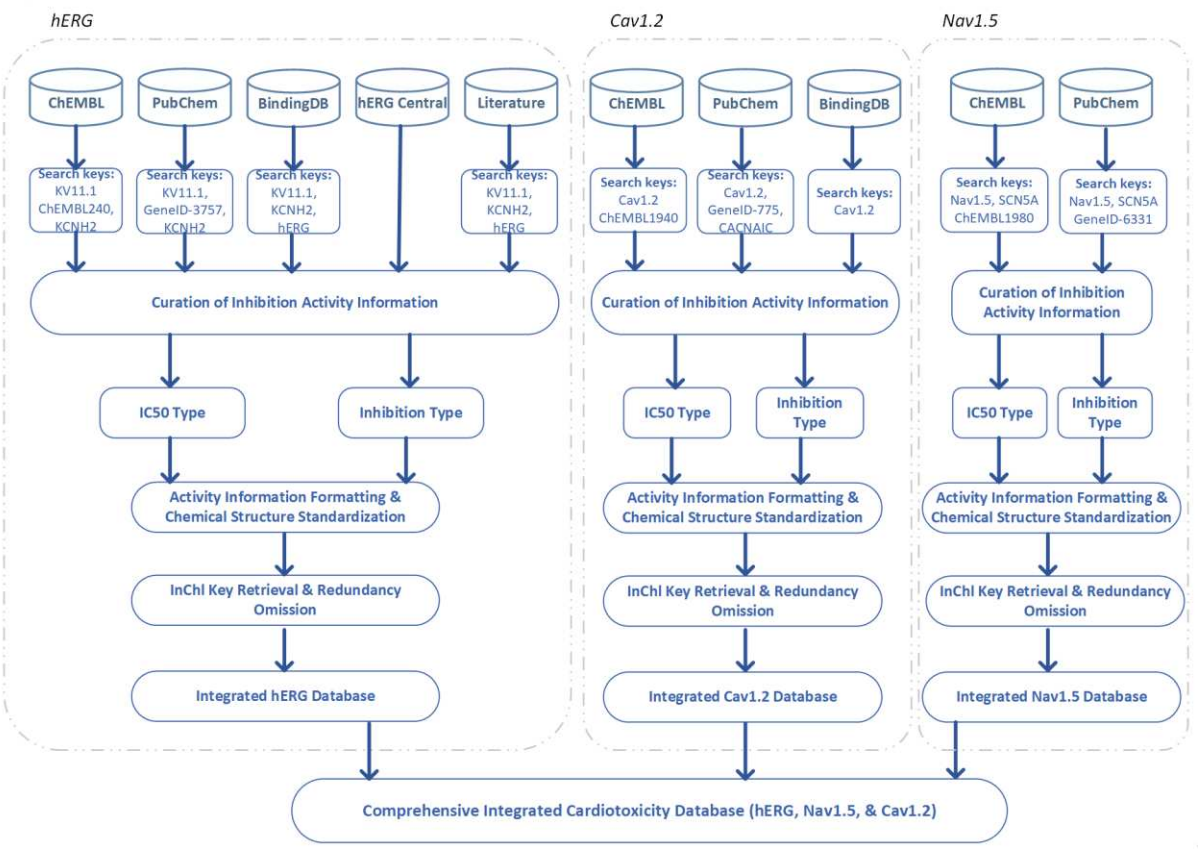

Figure 1. Schematic of the gathering, curation, and integration pipeline to generate the comprehensive integrated cardiotoxicity database of small molecules for all three ion channels: hERG, Nav1.5, and Cav1.2.

All activities were then transformed into molar values and subsequently standardized by calculating the pIC50 as follows :

$$pIC50 = -log_{10}(IC50) \quad .$$

Compounds were classified based on their IC50 values, following standard criteria used by researchers in the field [25][45][47]. Compounds with IC50 values of 10 µM or below (pIC50 ≥ 5) were categorized

as blockers (inhibitors), while compounds with IC50 values higher than 10 µM (pIC50 < 5) were categorized as non-blockers (inactive).

In adherence to data science best practices for model development, as illustrated in Figure 2, we extracted two external/independent test sets from each target dataset (hERG, Nav1.5, and Cav1.2). We used RDKit (default settings) to compute the Tanimoto similarity [48][49] between each pair of compounds in the datasets using 2048-bit extended connectivity fingerprints, also referred to as circular or Morgan fingerprints [50]. The first test set consisted of compounds with a structural similarity of no more than 60% (Tanimoto similarity ≤ 0.6) to the remaining development set, while the second test set comprised compounds with a structural similarity of no more than 70% (Tanimoto similarity ≤ 0.7) to the remaining development set. These external unique sets were denoted as hERG-70 & hERG-60 for hERG, Nav-70 & Nav-60 for Nav1.5, and Cav-70 & Cav-60 for Cav1.2. The composition of each set as derived from the upstream data sources can be observed in supplementary figure S1. Our unique development sets were subsequently partitioned into training and validation sets using an 80/20 ratio for each target. These splits were used for hyperparameter tuning and were stratified by pIC50. Note that random splitting of the training and validation sets can lead to splits that share portions of similar data points, which in turn, can lead to a positive bias in validation performance compared to our external test sets. However, any bias in validation performance is minimal due to multiple factors: the development set of curated molecular compounds that is already screened for distinct and unique structures, the stratification strategy, the size of the validation set, and the biological nature of our targets, namely, hERG, Nav1.5, and Cav1.2. These cardiac ion channels possess promiscuous binding sites at their central cavity that can interact with a diverse set of chemical structures [41][51][52]. In other words, hERG, Nav1.5, and Cav1.2 are not similar to an enzyme or a receptor with a specific catalytic site. Enzyme targets can usually interact with a distinct set of compounds with a common chemical scaffold. Given this additional biological parameter, which adds an intrinsic diversifying factor to our dataset, we believe that the derived validation and training datasets will already be diverse by nature.

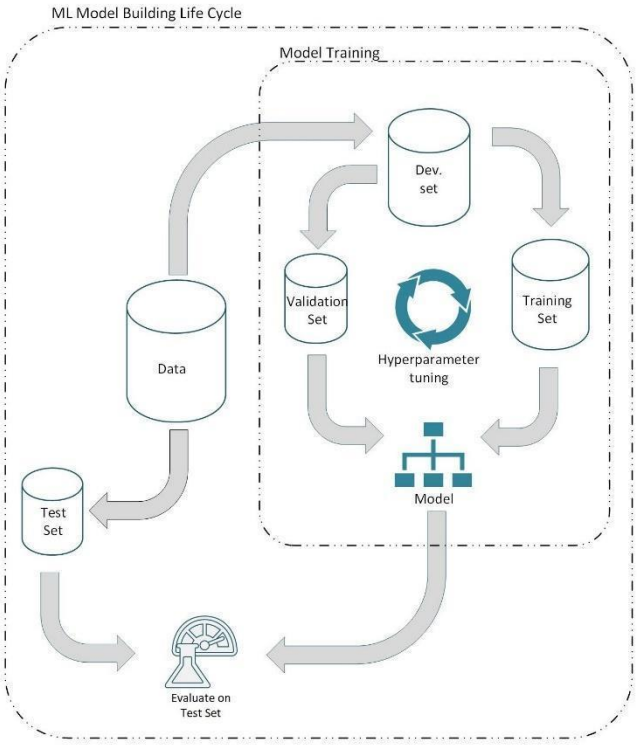

Figure 2. Data science methodology for hyperparameter tuning, monitoring overfitting, and best model selection.

2.2. Molecular Features

The chemical structures in our database are represented in the SMILES format and serve as the primary input for our predictive model training. Prior to conducting analyses, these structures need to be transformed into numerical representations. In this research, we explored three types of feature representations: molecular fingerprints, molecular descriptors, and molecular graph representations.

The PyBioMed Python package [53] was used to compute molecular fingerprints. This process generated two types of fingerprints: extended connectivity fingerprints with a maximum diameter parameter of two (ECFP2) (a vector of 1024 ECFP fingerprint values) and PubChem fingerprints (a vector of 881 values).

The calculation of molecular descriptors was performed using the Mordred Python package [54]. We employed 2D descriptors only as these require fewer computational resources compared to 3D descriptors without sacrificing predictive performance [14][55][56], resulting in a total of 1613 descriptors. Preprocessing and feature selection were accomplished through a Scikit-Learn [57] pipeline consisting of four modules. First, a univariate imputer was employed to discard columns with no calculated values and replace missing values in other columns with the mean. Second, a standardization step was applied to remove the mean and scale the values to unit variance. Third, zero-variance features were removed. Finally, for any pair of highly correlated features (Pearson correlation above 0.95), one of them was randomly discarded, as such correlated features convey nearly identical information. Consequently, the preprocessing procedure reduced the feature set for each target database (i.e. hERG, Nav1.15, and Cav1.2) to 806, 549, and 681 descriptors, respectively (refer to Tables S1, S2, and S3 of the supplementary data for the respective descriptor names).

Although SMILES are a convenient linear textual representation of molecules, a graph representation is closer to reality. Molecules can be represented as graphs, wherein atoms are depicted as nodes and

chemical bonds between atoms are depicted as edges. These two elements, nodes and edges, constitute the fundamental components of a graph and can serve as a molecular graph representation to encode the topological information inherent within molecules. To clarify more, this featurization process of a molecular compound yields two data structures. First, each atom/node is represented by a vector that can encode any type of information describing the atom. This is called the nodes embedding matrix, which has a dimension of M × k, with M the number of atoms in the molecule and k the number of node features. Here, we used 67 different node features according to the previously described node featurization strategy by Ryu et al. [59], including features such as the atom symbol, node degree, number of bound hydrogens, implicit valence, aromaticity, and size of the ring containing the atom, and further detailed in the supplementary table S4. As illustrated in supplementary Figures S2, S3, and S4, the value of M ranges up to 122 for hERG, 294 for Nav1.5, and 87 for Cav1.2. Second, the bonds/edges are encoded in an adjacency matrix that indicates the bonds between atoms [58]. As constrained by the deep learning model we used (see below for details), the adjacency matrix of size M × M is converted to an edge list of dimensions 2N × 2, with N the number of edges. RDKit was used to compute the node features and the edge list of the molecular structures.

2.3. Model Architecture

A neural network architecture was designed to accommodate the different features described above within a unified machine learning model. This system offers flexibility, allowing for the addition, removal, enabling, or disabling of different modules within the network. This flexibility is particularly advantageous during hyperparameter tuning. Figure 3 presents a simplified schematic of the architecture employed in this research.

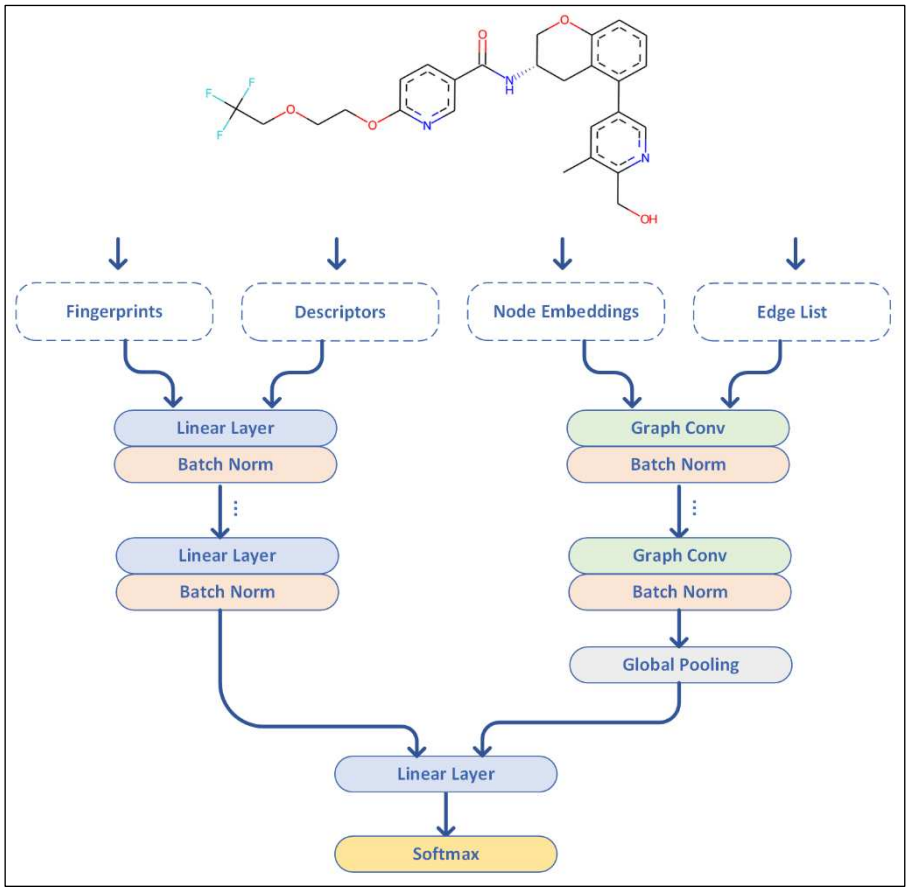

Figure 3. Simplified conceptual visualization of the deep learning model used in this study. The architecture can enable/disable any path, layer, or feature set of the molecule.

The architecture consists of two paths that are merged at the final layer. The first path accepts descriptors, fingerprints, or a combination of both feature vectors as input. The vector is then passed through a series of sequential blocks, with each block comprising a linear layer followed by batch normalization [60]. The second path takes the graph representation of the chemical structure as input. This graph representation is composed of a node embedding matrix with dimensions M × k and an edge list represented as a 2D matrix with dimensions 2N×2. These matrices undergo a sequence of feature extraction blocks, with each block consisting of a graph convolutional layer (GCN) [61] followed by batch normalization. For batch processing of chemical structures, zero padding was employed. Following the graph convolutions, a global pooling layer is applied to reduce the dimensionality of the generated tensor to a 1D vector. This vector is then concatenated with the intermediate representations that capture higher-level abstractions within the input chemical structure, as encoded by either the fingerprints or descriptors. The final output layer utilizes a softmax layer, predicting whether the compound is an inhibitor or not. Hyperparameter tuning was employed to find the best set of hyperparameters from a predefined dictionary as summarized in supplementary Table S5.

2.4. Evaluation Metrics

Initially, all models underwent evaluation on validation sets stratified by pIC50 derived from the development sets, consisting of 20% of the data. The best-performing hyperparameters for each combination of feature representations were identified based on their performance on the validation set. These hyperparameters were then used to construct the final best models, which were subsequently evaluated on external test sets comprising 60% and 70% structurally dissimilar molecular compounds .

The performance of the models was assessed using multiple binary evaluation metrics , including accuracy (AC), sensitivity (SE), specificity (SP), F1-score (F1), correct classification rate (CCR), and Matthew's correlation coefficient (MCC). Accuracy represents the overall predictive effectiveness of a classifier, while sensitivity and specificity measure the predictive powers for positive and negative instances, respectively. The CCR quantifies the proportion of instances that are correctly classified by the model. The F1-score computes the harmonic mean of precision and sensitivity. The MCC takes into account the balance ratios of the four categories in the confusion matrix (TP, TN, FP, FN) and provides an objective reflection of the model's predictive power. The final model selection was performed based on the F1 score.

The definitions of these evaluation metrics are provided as follows:

$$AC = \frac{TP+TN}{TP+FN+TN+FP}$$

$$SN = \frac{TP}{TP+FN}$$

$$PR = \frac{TP}{TP+FP}$$

$$SP = \frac{TN}{TN+FP}$$

$$F1 = 2 * \frac{SN*PR}{SN+PR}$$

$$CCR = \frac{SN+SP}{2}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP+FN) * (TP+FP) * (TN+FN) * (TN+FP)}}$$

2.5. hERG Benchmarked Tools Settings

2.5.1. CardioTox

CardioTox [21] is a command line-based prediction tool for hERG inhibition. The method uses an ensemble strategy of five models each trained on: 2D+3D descriptors, molecular graph representation, molecular fingerprints, SMILES embedding, and a fingerprint embedding model. The outputs of each model are concatenated, after which the final binary prediction (hERG blocker or non-blocker) is made. For benchmarking purposes we used the default settings as outlined in the corresponding GitHub repository (https://github.com/Abdulk084/CardioTox).

2.5.2. CardPred

CardPred [62] is a recent web-based prediction tool for hERG inhibition (http://ssbio.cau.ac.kr/CardPred). The method uses a deep learning network trained on a set of 2130 molecular compounds. The predictions are based on a combination of descriptors and fingerprints calculated using an external software, DRAGON (version 7.0.10) [63]. The tool predicts the probability of the chemical structure being a blocker or not. We used the default settings and a threshold of greater or equal than 50% to decide on hERG blockers in this study.

2.5.3. ADMETsar 2.0

ADMETsar 2.0 [64] is an optimized version of the previously released version of ADMETsar. It is a comprehensive open source and free tool for the prediction of different chemical ADMET properties. The web-based (http://lmmd.ecust.edu.cn/admetsar2/) prediction tool takes as input SMILES strings and comprises 47 different models for drug discovery, among them the hERG cardiotoxicity prediction model. Default settings were used to benchmark this tool.

2.5.4. ADMETlab 2.0

ADMETlab 2.0 [65] is another comprehensive web server (https://admetmesh.scbdd.com/) for the predictions of pharmacokinetics and toxicity properties of chemicals using a multi-task graph attention framework. The web-based prediction tool takes as input SMILES strings to make multiple predictions, including hERG cardiotoxicity predictions. Default settings were used to benchmark this tool.

3. Results

3.1. A Comprehensive Database of Cardiac Ion Channel Blockers

The presented collection of data establishes a framework intended for researchers operating within the realm of drug discovery to conduct in-depth analyses and further studies. This collection includes a large and freely accessible unique and comprehensive hERG, Nav1.5, and Cav1.2 cardiotoxicity integrated database of small molecules and their activities. The database was sourced from a variety of public repositories, such as databases like ChEMBL, PubChem, BindingDB, and hERGCentral, as well as from US patent data and literature mining. Figure 4 shows a quick overview of the data composition for each target (hERG, Nav1.5, and Cav1.2), indicating the sources from which it was gathered.

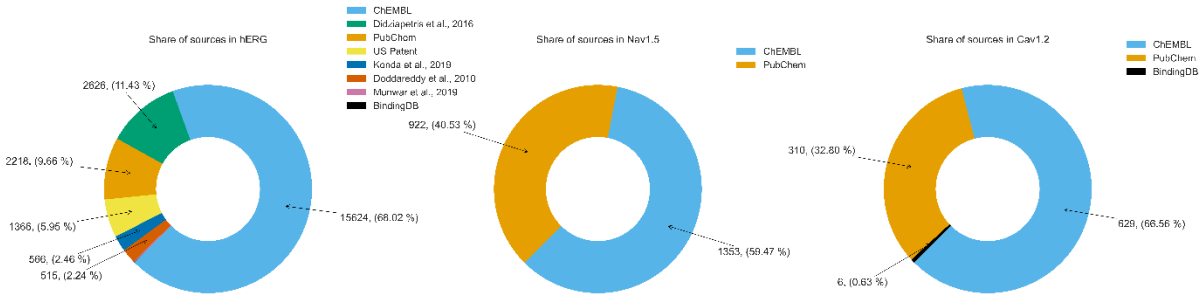

Figure 4: The composition of each target (hERG, Nav1.5, and Cav1.2) as derived from the upstream data sources in the final extensive cardiotoxicity database.

External test sets were derived from the collected database using two Tanimoto similarity thresholds, namely 60% and 70% structural similarity, as illustrated in the pairwise Tanimoto similarity distributions in Figure 5. These sets were extracted in a manner that ensures the preservation of the pIC50 distribution found in the development set. This approach ensures that the blocker vs. non-blocker class distribution is maintained within each set, thereby enabling reliable performance evaluation of the built models. Figure 6 illustrates the pIC50 density distribution in each dataset, which also supports the community choice of 10µM as a threshold for discriminating between blockers and non-blockers. In order to verify that the training and test datasets are independent and identically distributed (i.i.d.), we utilized t-SNE [76], a dimensionality reduction technique. The 1905 fingerprints, explained in section 2.2, were employed as input for the t-SNE process. As depicted in Figure 7, the chemical space shows a strong overlap between the training and test compounds, indicating that the i.i.d. assumption is satisfied, which is essential for the proper development of machine learning models (supplementary Figure S14, Figure S15, and Figure S16 further demonstrate this based on physicochemical properties). The distribution of compounds in each class exhibits a ratio of approximately 6:4, favoring blockers, for each respective set, as illustrated in Table 1.

Table 1. Class distribution of blockers vs. non-blockers in each set and for each target ion channel.

	hERG		**Nav1.5**		**Cav1.2**	
Property	Blockers	Non-Blockers	Blockers	Non-Blockers	Blockers	Non-Blockers
Dev. Set	13428	8818	1414	655	530	272
Eval-70	264	209	97	45	52	29
Eval-60	125	125	48	16	41	21

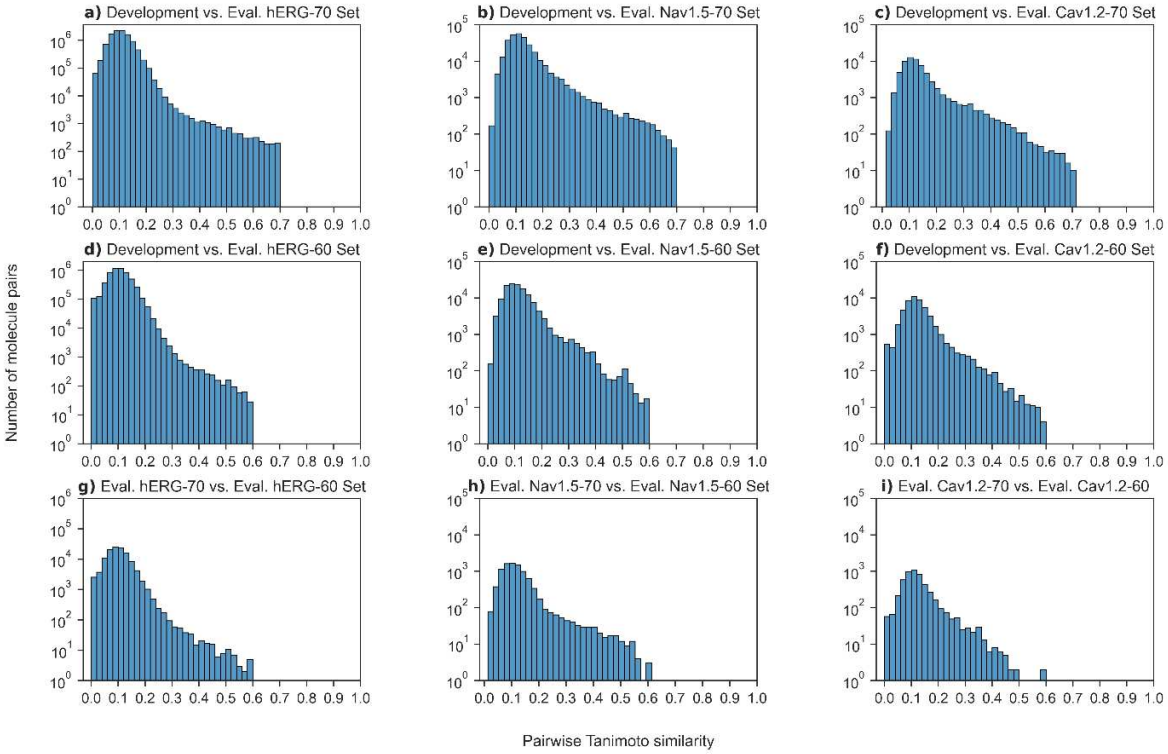

Figure 5. Distribution of the pairwise Tanimoto similarity for each molecule in the **(a)** hERG development set with the ones in the evaluation set hERG-70, **(b)** Nav1.5 development set with the ones in the evaluation set Nav-70, **(c)** Cav1.2 development set with the ones in the evaluation set Cav-70, **(d)** hERG development set with the ones in the evaluation set hERG-60, **(e)** Nav1.5 development set with the ones in the evaluation set Nav-60, **(f)** Cav1.2 development set with the ones in the evaluation set Cav-60, **(g)** hERG-70 evaluation set with the ones in the evaluation set hERG-60, **(h)** Nav-70 evaluation set with the ones in the evaluation set Nav-60, and **(i)** the Cav-70 evaluation set with the ones in the evaluation set Cav-60.

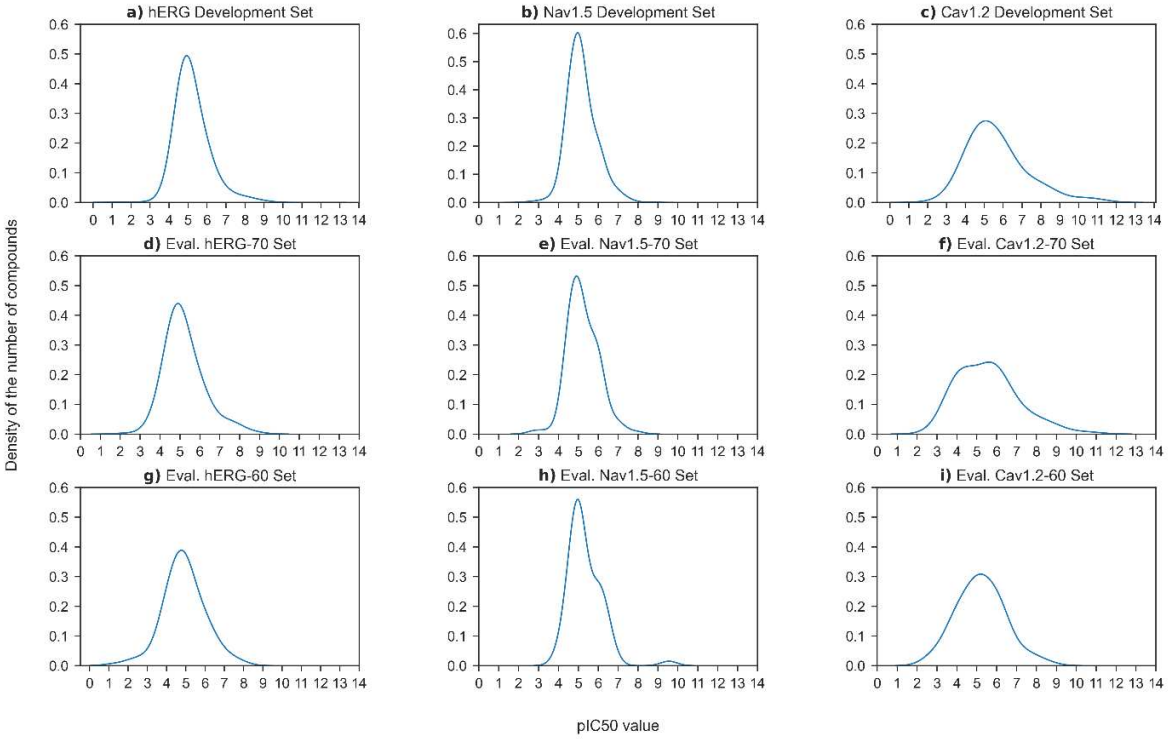

Figure 6. Density distribution of the pIC50 activity of molecular compounds in the **(a)** hERG development set, **(b)** Nav1.5 development set, **(c)** Cav1.2 development set, **(d)** hERG-70 evaluation set, **(e)** Nav-70 evaluation set, **(f)** Cav-70 evaluation set, **(g)** hERG-60 evaluation set, **(h)** Nav-60 evaluation set, and **(i)** Cav-60 evaluation set.

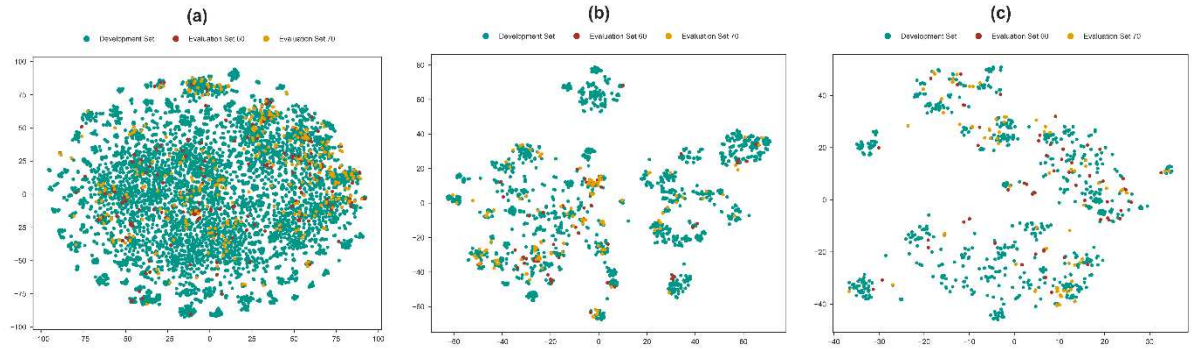

Figure 7: t-SNE visualizations showing the overlap between the development and the two external test sets (Eval-60 and Eval-70) for (a) hERG (b) Nav1.5 and (c) Cav1.2.

3.2. Evaluation of Molecular Features For Machine Learning of Cardiotoxicity

The deep learning model presented in this study has been designed to take different combinations of feature representations of a molecular structure as input. For each target's (i.e. hERG, Nav1.5, Cav1.2) molecular compounds dataset, the network was trained on seven combinations from the three feature sets (fingerprint, descriptor, and graph representation). Conducting a thorough exploration of the hyperparameter search space, as detailed in supplementary Table S5, yields a total of 576 distinct combinations for each feature combination. In order to mitigate the computational cost associated with this exhaustive procedure, a random search policy [77] was employed, involving only 100 randomly selected hyperparameter combinations. The final optimal set of hyperparameters was determined based on optimal performance on the validation set, which accounts for 20% of the

development set. The training process was performed on an NVIDIA T4 with 16GB dedicated RAM, granted by Google Cloud Research. For each feature combination, the best-performing model was selected and subsequently evaluated on our two external datasets. We name the best models CToxPred-hERG, CToxPred-Nav, and CToxPred-Cav for each of the targets respectively.

3.2.1. hERG Cardiotoxicity Prediction

We assessed the optimal model for each combination of molecular features on two separate external test sets to evaluate hERG cardiotoxicity predictions. Surprisingly, the results revealed that standard fingerprints outperformed more complex feature representations. Nonetheless, it is worth noting that the combination of fingerprints with additional complex features showcased competitive performance, potentially enhancing generalization. This trend was demonstrated by the competitive performance of the combined features on hERG-60 versus hERG-70 datasets. However, the more different the test data is from the training data, the lower the performance, as evidenced by the drop in performance from the hERG-70 to hERG-60 test sets (an observation that was expected). For additional observations and insights, please consult Table 2 and supplementary Figure S8.

Considering all evaluation metrics, we can rank the individual feature sets in terms of performance from best to worst as follows: fingerprints, descriptors, and then graph representations. The fingerprint-based model exhibited the highest accuracy and F1-score for both evaluation sets. The F1-score for hERG-60 demonstrated a comparative value to the models that utilized all combined features. Therefore, we have selected the fingerprint-based model as our best-performing model for predicting hERG cardiotoxicity of small molecules (see supplementary Table S6 for a bootstrapping analysis of the robustness of the models for different feature sets). Throughout the rest of the analysis, we will refer to this model as CToxPred-hERG. For the confusion matrices of CToxPred-hERG on the two external test sets, please refer to supplementary Figure S9.

Table 2. hERG toxicity prediction performance of each feature combination using the deep learning model on the two external test sets hERG-70 and hERG-60.

	hERG-70						**hERG-60**					
Model Input Features	AC	F1	SN	SP	CCR	MCC	AC	F1	SN	SP	CCR	MCC
Fingerprints	**81.4**	**83.9**	86.7	**74.6**	**80.7**	**62.1**	**71.2**	72.9	77.6	**64.8**	**71.2**	**42.8**
Descriptors	77.6	81.1	86.0	67.0	76.5	54.4	66.8	70.0	77.6	56.0	66.8	34.4
Graph	74.8	80.4	**92.4**	52.6	72.5	50.1	64.0	70.6	86.4	41.6	64.0	31.3
Fingerprints+ Descriptors	78.2	80.8	82.2	73.2	77.7	55.7	69.2	71.4	76.8	61.6	69.2	38.9
Fingerprints+ Graph	79.1	82.5	88.3	67.5	77.9	57.5	66.9	70.3	78.4	55.2	66.8	34.5
Descriptors+ Graph	76.7	79.8	82.2	69.9	76.0	52.6	66.0	66.9	68.8	63.2	66.0	32.1
All Combined	73.2	78.2	86.4	56.5	71.4	45.4	68.4	**73.8**	**88.8**	48.0	68.4	40.3

We also conducted a comprehensive benchmarking study of CToxPred-hERG, comparing it with another published command-line tool called CardioTox [21], as well as three web-based prediction tools: CardPred [62], ADMETsar 2.0 [64], and ADMETlab 2.0 [65]. All models were evaluated using the same external sets of molecular compounds (Table 3). As anticipated, all models demonstrated superior performance on the hERG-70 dataset and relatively lower performance on the hERG-60 dataset.

Among the competing models, CardPred exhibited the lowest performance, while CardioTox and CToxPred-hERG displayed comparable results on the hERG-70 dataset. However, CardioTox outperformed our best model on the hERG-60 dataset. Two hypotheses can be inferred from these findings: either CardioTox can generalize better to unknown molecules, or the evaluation data used for CardioTox included compounds that were already present in its training data. Further investigations revealed that the training data used to construct the CardioTox model had a 40% overlap with the hERG-70 dataset and a 68% overlap with the hERG-60 dataset. As such, we anticipate that there is a positive bias in these evaluation results for CardioTox. In order to conduct an unbiased comparison between the two models, it would be necessary for both methods to be trained on the same data and evaluated using the same test set.

Table 3. Performance evaluation of CToxPred-hERG compared to CardPred, ADMETsar, ADMETlab, and CardioTox on the 2 external test sets.

	hERG-70						**hERG-60**					
Model	AC	F1	SN	SP	CCR	MCC	AC	F1	SN	SP	CCR	MCC
CardPred	56.1	57.0	52.7	60.3	56.5	13.0	53.9	45.4	37.9	70.2	54.1	8.6
ADMETsar 2.0	68.5	75.0	84.5	48.3	66.4	35.5	66.4	70.6	80.8	52.0	66.4	34.3
ADMETlab 2.0	71.7	73.8	71.6	71.8	71.7	43.1	68.0	67.5	66.4	69.6	68.0	36.0
CardioTox	81.2	83.1	83.0	**78.9**	**81.0**	61.9	**80.4**	**78.4**	71.2	**89.6**	**80.4**	**61.9**
CToxPred-hERG	**81.4**	**83.9**	**86.7**	74.6	80.7	**62.1**	71.2	72.9	**77.6**	64.8	71.2	42.8

3.2.2. Nav1.5 Cardiotoxicity Prediction

Similar to the evaluation performed for hERG, we also assessed the most effective predictive model for Nav1.5 cardiotoxicity using various combinations of molecular features on the two external test sets. For the Nav-70 test set, three feature combinations demonstrated superior and comparable performance in terms of both F1-score and AC. These combinations were *'fingerprints + descriptors'*, *'fingerprints + graph representation'*, and *'all combined'* features. Thus, we can infer that, similar to hERG, fingerprints provide more informative representations to discriminate blocksers from non-blockers. This conclusion is further supported by the superior performance exhibited by all features in the hERG-70 set. However, it is worth noting that this observation does not hold for structurally dissimilar compounds, as demonstrated in the Nav-60 evaluation, where descriptors performed better (as shown in Table 4 and supplementary Figure S10).

Considering different evaluation metrics and individual feature sets, we can rank the best-performing features from worst to best as follows: fingerprints, graphs, and then molecular descriptor representations. Although the ranking order may vary when considering each test set individually, we tend to prioritize the Nav-60 set as it is structurally dissimilar and therefore represents better generalization to unseen data. Interestingly, the combination of fingerprints and descriptors exhibited the highest performance across most metrics and on both test sets. This combination can be seen as an enrichment of complementary feature information, where one set excels at predicting Nav-70 compounds while the other performs better for Nav-60 molecular structures.

Based on these findings, the model that combines fingerprints and descriptors was selected as the optimal model for the Nav1.5 cardiotoxicity prediction task and named CToxPred-Nav1.5 (see supplementary Table S6 for a bootstrapping analysis of the robustness of the models for different feature sets). For the confusion matrices detailing the predictions of CToxPred-Nav1.5 on the two external test sets, please refer to the supplementary Figure S11.

Table 4. Nav1.5 toxicity prediction performance of each feature combination using the deep learning model on the two external test sets Nav-70 and Nav-60.

	Nav-70						**Nav-60**					
Model Input Features	AC	F1	SN	SP	CCR	MCC	AC	F1	SN	SP	CCR	MCC
Fingerprints	80.3	85.3	83.5	73.3	78.4	55.6	62.5	72.7	66.7	50.0	58.3	14.9
Descriptors	79.6	84.7	82.5	73.3	77.9	54.4	75.0	83.0	81.2	56.2	68.8	36.1
Graph	77.5	83.7	84.5	62.2	73.4	47.3	70.3	78.7	72.9	**62.5**	67.7	32.0
Fingerprints + Descriptors	**81.7**	**86.5**	**85.6**	73.3	79.5	58.2	**76.6**	**84.2**	**83.3**	56.2	**69.8**	**38.8**
Fingerprints + Graph	**81.7**	**86.5**	**85.6**	73.3	79.5	58.2	60.9	72.5	68.8	37.5	53.1	5.8
Descriptors + Graph	78.2	83.4	80.4	73.3	76.9	51.9	71.9	80.9	79.2	50.0	64.6	28.1
All Combined	**81.7**	86.2	83.5	**77.8**	**80.6**	**59.4**	71.9	81.6	**83.3**	37.5	60.4	21.8

3.2.3. Cav1.2 Cardiotoxicity Prediction

Similarly to hERG and Nav1.5, we conducted an evaluation of the most effective predictive model for Cav1.2 cardiotoxicity, using various combinations of molecular features, on two external test sets. One notable observation, distinguishing Cav1.2 from other targets, is the significant disparity in predictive performance between the Cav-70 and Cav-60 test sets (as depicted in Table 5 and supplementary Figure S12). This discrepancy can be attributed to the size of the available data we could gather for training, as compared to Nav1.5 and hERG, the training dataset for Cav1.2 was considerably smaller.

Notwithstanding the data limitations, the models exhibited impressive results for the Cav-70 test set across all feature combinations, achieving an average accuracy of approximately 82% and an F1-score above 85%. This suggests that the task of predicting Cav1.2 cardiotoxicity is relatively easy compared

to the previous two targets. When considering individual feature sets and all evaluation metrics, we can rank the best-performing features from worst to best as follows: graph representation is the weakest, followed by fingerprints, and then molecular descriptor representations. Descriptors demonstrate high informative information in this task compared to the other two features. When combined with fingerprints, the performance is significantly enhanced. Similar to Nav1.5, the optimal model for Cav1.2 combines both fingerprint and descriptor-based representations (see supplementary Table S6 for a bootstrapping analysis of the robustness of the models for different feature sets). We have named this model CToxPred-Cav1.2. For the confusion matrices detailing the predictions of CToxPred-Cav1.2 on the two external test sets, please refer to supplementary Figure S13.

Table 5. Cav1.2 toxicity prediction performance of each feature combination using the deep learning model on the two external test sets Cav-70 and Cav-60.

	Cav-70						**Cav-60**					
Model Input Features	AC	F1	SN	SP	CCR	MCC	AC	F1	SN	SP	CCR	MCC
Fingerprints	82.7	87.5	94.2	62.1	78.1	61.6	59.7	70.6	**73.2**	33.3	53.3	6.8
Descriptors	85.2	89.1	94.2	69.0	81.6	67.2	64.5	71.8	68.3	57.1	62.7	24.5
Graph	75.3	81.5	84.6	58.6	71.6	44.9	59.7	66.7	61.0	57.1	59.1	17.2
Fingerprints + Descriptors	**86.4**	**90.1**	96.2	69.0	82.6	**70.2**	**69.4**	**75.9**	**73.2**	**61.9**	**67.5**	**34.1**
Fingerprints + Graph	84.0	88.7	**98.1**	58.6	78.3	65.4	59.7	69.1	68.3	42.9	55.6	11.0
Descriptors + Graph	80.2	85.2	88.5	65.5	77.0	56.0	62.9	70.9	68.3	52.4	60.3	20.1
All Combined	**86.4**	89.9	94.2	**72.4**	**83.3**	70.0	64.5	71.8	68.3	57.1	62.7	24.5

4. Conclusion

Cardiac voltage-gated ion channels are collectively responsible for generating the action potential, which is required for cardiac cells' contraction. Notably, the hERG, Nav1.5, and Cav1.2 ion channels are key components of the cardiac action potential, and their inhibition by drugs can cause severe cardiovascular complications. Therefore, accurate prediction of the potential cardiotoxic liability of these channels in drug interactions is crucial. This research addresses this need by assembling a large and comprehensive dataset of small molecules specifically tailored for this purpose. An important element of our effort is that the cardiotoxicity database is freely available as open access for further community development of machine cardiotoxicity prediction models. This is in contrast to previous efforts that have employed proprietary and private datasets that are not publicly accessible for the scientific community, such as GOSTAR [45].

Standard datasets are necessary for the proper comparison of different tools. We seek through this study to emphasize the importance of establishing a robust and comprehensive framework consisting of extensive and publicly available development and external test sets. Such a framework enables the scientific community to focus on developing AI models while facilitating the straightforward benchmarking of model performance; Hence avoiding over-optimistic results as presented earlier in the introduction in Konda et al. [26] and Doddareddy et al. [28] data as well as in the data leakage by CardioTox [21] observed in the sub-section 3.2.1 of the results section. Furthermore, in order to assess the generalizability of developed tools, it is necessary to evaluate on structurally dissimilar datasets. The more different the test data is from the training data, the lower the performance, as evidenced by the drop in performance from the Eval-70 to Eval-60 test sets for each of the targets.

Additionally, a deep learning model has been developed that uses multiple types of feature representations, including fingerprints, descriptors, and molecular graph representations. A thorough benchmarking of these representations has been conducted to identify the most effective combinations for each respective task. Overall, the results demonstrate that descriptors possess higher predictive power, resulting in improved models and enhanced generalizability, particularly for Nav1.5 and Cav1.2. The results have confirmed a finding by Jiang et al. [78], who did empirically demonstrate that on average the descriptor-based models outperform the graph-based models in terms of prediction accuracy and computational efficiency. In the case of hERG, fingerprints alone prove sufficient for discriminating blockers from non-blockers. However, for Nav1.5 and Cav1.2, the combination of fingerprints and descriptors yields the best performance. These results demonstrate how simple features can perform better than more complex ones, such as GNN-based features. As a result, we advocate for carefully evaluating various feature representations instead of immediately using the most complex deep learning models possible.

The database utilized in this study is intended to serve as a comprehensive framework for researchers in the field, enabling them to build predictive models and easily benchmark their results using consistent test sets. It is publicly available as open access on Zenodo at https://zenodo.org/record/8359714. As a result of this research, robust models for predicting cardiotoxicity for each ion channel have been developed and consolidated into a comprehensive tool named CToxPred.

In terms of future work, the collected dataset could be leveraged to develop regression models capable of directly predicting the estimated potency of molecular compounds. Additionally, structural modeling could be employed to validate the results. Other potential projects may explore the application of data augmentation techniques, particularly for Cav1.2 and Nav1.5, to assess improvements in predictions.

Acknowledgement

The authors would like to thank Chanuka Fernando for his participation during the data collection.

Data and Software Availability

CToxPred, a comprehensive cardiotoxicity prediction method, is available as an open-source Python command-line tool and can be called from a notebook. It uses Pybel (version 0.13.2) [66], Open Babel (version 3.1.1) [67], and PyBioMed (version 1.0) [53] to compute PubChem & ECFP2 fingerprints; Mordred (version 1.2.0) [54] to calculate molecular descriptors; RDKit (version 2022.09.1) [68] for chemical structure information retrieval, used for example in the graph representation construction

and other tasks; Scikit-Learn (version 1.0.2) [57] for pipeline data preprocessing and evaluation metric calculations; PyTorch (version 1.12.1) & PyTorch Geometric (version 2.3.1) [69] libraries to implement our deep learning and graph neural network architecture; and NumPy (version 1.21.6) [70], SciPy (version 1.7.3) [71], and Pandas (version 1.3.5) [72] for scientific computing. Matplotlib (version 3.5.3) [73] and Seaborn (version 0.12.2) [74] were used for data visualization. Data analysis was performed using Jupyter notebooks [75].

CToxPred is available as open source under the permissive MIT license on GitHub at https://github.com/issararab/CToxPred . Analysis notebooks to reproduce the presented results are also available in the same repository at https://github.com/issararab/CToxPred/blob/main/notebooks/ .

All data used in this study is freely available through Zenodo for long-term archival at https://zenodo.org/record/8359714 . The data is also available on the CToxPred GitHub repository, alongside precomputed molecular fingerprints and descriptors for faster reproducibility of the results at https://github.com/issararab/CToxPred/tree/main/data .

Supporting Information

Figure S1: The composition of each set as derived from the upstream data sources in the final extensive cardiotoxicity database. **Figure S2:** Atom composition analysis of molecules in our hERG development set. **Figure S3:** Atom composition analysis of molecules in our Nav1.5 development set. **Figure S4:** Atom composition analysis of molecules in our Cav1.2 development set. **Figure S5:** Distributions of the 8 physicochemical properties between inhibitor (blocker) and inactive (non-blocker) compounds in the hERG dataset. **Figure S6:** Distributions of the 8 physicochemical properties between inhibitor (blocker) and inactive (non-blocker) compounds in the Nav1.5 dataset. **Figure S7:** Distributions of the 8 physicochemical properties between inhibitor (blocker) and inactive (non-blocker) compounds in the Cav1.2 dataset. **Figure S8:** Accuracy performance comparison of different feature combinations on hERG liability prediction. **Figure S9:** CToxPred-hERG confusion matrixes. **Figure S10:** Accuracy performance comparison of different feature combinations on Nav1.5 liability prediction. **Figure S11:** CToxPred-Nav1.5 confusion matrixes. **Figure S12:** Accuracy performance comparison of different feature combinations on Cav1.2 liability prediction. **Figure S13:** CToxPred-Cav1.2 confusion matrixes. **Figure S14:** Distributions of physicochemical properties between the development and the test sets of compounds in the hERG dataset. **Figure S15:** Distributions of physicochemical properties between the development and the test sets of compounds in the Nav1.5 dataset. **Figure S16:** Distributions of physicochemical properties between the development and the test sets of compounds in the Cav1.2 dataset. **Table S1.** List of hERG molecular descriptors used in the development of the descriptor-based predictive model. **Table S2.** List of Nav1.5 molecular descriptors used in the development of the descriptor-based predictive model. **Table S3.** List of Cav1.2 molecular descriptors used in the development of the descriptor-based predictive model. **Table S4**. Information on atom descriptors used as node features for the development of the graph-based GCN module. **Table S5.** Hyperparameters considered in the study. **Table S6.** Bootstrapping analysis of the prediction models on the test sets. This material is available free of charge via the Internet at http://pubs.acs.org.

References

[1]. Oprea, T. I.; & Matter, H. Integrating virtual screening in lead discovery. Current opinion in chemical biology, **2004**, *8*(4), 349-358.

[2]. Dean, A.; Lewis, S. (Eds.). Screening: methods for experimentation in industry, drug discovery, and genetics. Springer Science & Business Media. **2006**

[3]. Bailey, J.; Balls, M. Recent efforts to elucidate the scientific validity of animal-based drug tests by the pharmaceutical industry, pro-testing lobby groups, and animal welfare organisations. BMC Medical Ethics, **2019**, *20*(1), 1-7.

[4]. DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of health economics, **2016**, *47*, 20-33.

[5]. Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery, **2010**, *9*(3), 203-214.

[6]. Pu, L.; Naderi, M.; Liu, T.; Wu, H. C.; Mukhopadhyay, S.; Brylinski, M. etoxpred: A machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacology and Toxicology, **2019**, *20*, 1-15.

[7]. Raies, A. B.; Bajic, V. B. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdisciplinary Reviews: Computational Molecular Science, **2016**, *6*(2), 147-172.

[8]. Darpo, B.; Nebout, T.; Sager, P. T. Clinical evaluation of QT/QTc prolongation and proarrhythmic potential for nonantiarrhythmic drugs: the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use E14 guideline. The Journal of Clinical Pharmacology, **2006**, *46*(5), 498-507.

[9]. Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of clinical drug development fails and how to improve it?. Acta Pharmaceutica Sinica B, **2022**, *12*(7), 3049-3062.

[10]. Villoutreix, B. O. ; Taboureau, O. Computational investigations of hERG channel blockers: New insights and current predictive models. Advanced drug delivery reviews, **2015**, *86*, 72-82.

[11]. Brown, A. M. Drugs, hERG and sudden death. Cell calcium, **2004**, 35(6), 543-547.

[12]. Aronov, A. M. Predictive in silico modeling for hERG channel blockers. Drug discovery today, **2005**, *10*(2), 149-155.

[13]. Hung, C. L.; Chen, C. C. Computational approaches for drug discovery. Drug development research, **2014**, *75*(6), 412-418.

[14]. Arab, I.; Barakat, K. ToxTree: descriptor-based machine learning models for both hERG and Nav1.5 cardiotoxicity liability predictions. *arXiv*, December 27, 2021, arXiv:2112.13467, ver. 1. DOI: 10.48550/arXiv:2112.13467v1

[15]. Wang, S.; Li, Y.; Xu, L.; Li, D.; Hou, T. Recent developments in computational prediction of HERG blockage. Current topics in medicinal chemistry, **2013**, *13*(11), 1317-1326.

[16]. Sun, H. An accurate and interpretable Bayesian classification model for prediction of hERG liability. ChemMedChem: Chemistry Enabling Drug Discovery, **2006**, *1*(3), 315-322.

[17]. Obrezanova, O.; Csányi, G.; Gola, J. M.; Segall, M. D. Gaussian processes: a method for automatic QSAR modeling of ADME properties. Journal of chemical information and modeling, **2007**, *47*(5), 1847-1857.

[18]. Kar, S.; Roy, K. Prediction of hERG potassium channel blocking actions using combination of classification and regression based models: a mixed descriptors approach. Molecular Informatics, **2012**, *31*(11-12), 879-894.

[19]. Broccatelli, F.; Mannhold, R.; Moriconi, A.; Giuli, S.; Carosati, E. QSAR modeling and data mining link Torsades de Pointes risk to the interplay of extent of metabolism, active transport, and HERG liability. Molecular pharmaceutics, **2012**, *9*(8), 2290-2301.

[20]. Perry, M.; Sanguinetti, M.; Mitcheson, J. Symposium review: revealing the structural basis of action of hERG potassium channel activators and blockers. The Journal of physiology, **2010**, *588*(17), 3157-3167.

[21]. Karim, A.; Lee, M.; Balle, T.; Sattar, A. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. Journal of Cheminformatics, **2021**, *13*(1), 1-13.

[22]. Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep learning-based prediction of drug-induced cardiotoxicity. Journal of chemical information and modeling, **2019**, *59*(3), 1073-1084.

[23]. Karim, A.; Riahi, V.; Mishra, A.; Newton, M.H.; Dehzangi, A.; Balle, T.; Sattar, A. Quantitative toxicity prediction via meta ensembling of multitask deep learning models. Acs Omega, **2021**, *6*(18), 12306-12317.

[24]. Sarkar, A.; Bhavsar, A. Virtual Screening of Pharmaceutical Compounds with hERG Inhibitory Activity (Cardiotoxicity) using Ensemble Learning. *arXiv*, June 5, 2021, arXiv:2106.04377, ver. 1. DOI:10.48550/arXiv.2106.04377

[25]. Tran, T. T. V.; Surya Wibowo, A.; Tayara, H.; Chong, K. T. Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives. Journal of Chemical Information and Modeling, **2023**, *63*(9), 2628-2643.

[26]. Konda, L. S. K.; Praba, S. K.; Kristam, R. hERG liability classification models using machine learning techniques. Computational Toxicology, **2019**, *12*, 100089.

[27]. Liu, L. L.; Lu, J.; Lu, Y.; Zheng, M. Y.; Luo, X. M.; Zhu, W. L.; Jiang, H.L.; Chen, K. X. Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. Acta Pharmacologica Sinica, **2014**, *35*(8), 1093-1102.

[28]. Doddareddy, M. R.; Klaasse, E. C.; Shagufta, IJzerman, A. P.; Bender, A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. ChemMedChem, **2010**, *5*(5), 716-729.

[29]. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J.P. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research, **2012**, *40*(D1), D1100-D1107.

[30]. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M. The ChEMBL bioactivity database: an update. Nucleic acids research, **2014**, *42*(D1), 1083-1090.

[31]. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; Davies, M. The ChEMBL database in 2017. Nucleic acids research, **2017**, *45*(D1), 945-954.

[32]. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; Zaslavsky, L. PubChem in 2021: new data content and improved web interfaces. Nucleic acids research, **2021**, *49*(D1), 1388-1395.

[33]. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research, **2007**, *35*(suppl_1), D198-D201.

[34]. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, **2016**, *44*(D1), D1045-D1053.

[35]. Du, F.; Yu, H.; Zou, B.; Babcock, J.; Long, S.; Li, M. hERGCentral: a large database to store, retrieve, and analyze compound-human Ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. Assay and drug development technologies, **2011**, *9*(6), 580-588.

[36]. Didziapetris, R.; Lanevskij, K. Compilation and physicochemical classification analysis of a diverse hERG inhibition database. Journal of computer-aided molecular design, **2016**, *30*, 1175-1188.

[37]. Konda, L. S. K.; Praba, S. K.; Kristam, R. hERG liability classification models using machine learning techniques. Computational Toxicology, **2019**, *12*, 100089.

[38]. Doddareddy, M. R.; Klaasse, E. C.; Shagufta, IJzerman, A. P.; Bender, A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. ChemMedChem, **2010**, *5*(5), 716-729.

[39]. Munawar, S.; Vandenberg, J. I.; Jabeen, I. Molecular docking guided grid-independent descriptor analysis to probe the impact of water molecules on conformational changes of hERG inhibitors in drug trapping phenomenon. International Journal of Molecular Sciences, **2019**, *20*(14), 3385.

[40]. AlRawashdeh, S.; Chandrasekaran, S.; Barakat, K. H. Structural analysis of hERG channel blockers and the implications for drug design. Journal of Molecular Graphics and Modelling, **2023**, *120*, 108405.

[41]. Kalyaanamoorthy, S.; Barakat, K. H. Development of safe drugs: the hERG challenge. Medicinal research reviews, **2018**, *38*(2), 525-555.

[42]. Guo, L.; Guthrie, H. Automated electrophysiology in the preclinical evaluation of drugs for potential QT prolongation. Journal of pharmacological and toxicological methods, **2005**, *52*(1), 123-135.

[43]. Dimmitt, S.; Stampfer, H.; Martin, J. H. When less is more–efficacy with less toxicity at the ED50. British Journal of Clinical Pharmacology, **2017**, *83*(7), 1365.

[44]. Di Veroli, G. Y.; Davies, M. R.; Zhang, H.; Abi-Gerges, N.; Boyett, M. R. High-throughput screening of drug-binding dynamics to HERG improves early drug safety assessment. American Journal of Physiology-Heart and Circulatory Physiology, **2013**, *304*(1), H104-H117.

[45]. Sato, T.; Yuki, H.; Ogura, K.; Honma, T. Construction of an integrated database for hERG blocking small molecules. PLoS One, **2018**, *13*(7), e0199348.

[46]. O'Boyle, N. M. Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. Journal of cheminformatics, **2012**, *4*, 1-14.

[47]. Zhang, X.; Mao, J.; Wei, M.; Qi, Y.; Zhang, J. Z. Hergspred: Accurate classification of herg blockers/nonblockers with machine-learning models. Journal of chemical information and modeling, **2022**, *62*(8), 1830-1839.

[48]. Willett, P. Similarity-based virtual screening using 2D fingerprints. Drug discovery today, **2006**, *11*(23-24), 1046-1053.

[49]. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, **2015**, *7*(1), 1-13.

[50]. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. Journal of chemical information and modeling, **2010**, *50*(5), 742-754.

[51]. Butler, A.; Helliwell, M. V.; Zhang, Y.; Hancox, J. C.; Dempsey, C. E. An update on the structure of hERG. Frontiers in pharmacology, **2020**, *10*, 1572.

[52]. Li, Z.; Jin, X.; Wu, T.; Huang, G.; Wu, K.; Lei, J.; Pan, X.; Yan, N. Structural basis for pore blockade of the human cardiac sodium channel Nav1. 5 by the antiarrhythmic drug quinidine. Angewandte Chemie, **2021**, *133*(20), 11575-11581.

[53]. Dong, J.; Yao, Z.J.; Zhang, L.; Luo, F.; Lin, Q.; Lu, A.P.; Chen, A.F.; Cao, D.S. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. Journal of cheminformatics, **2018**, *10*, 1-11.

[54]. Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. Journal of cheminformatics, **2018**, *10*(1), 1-14.

[55]. Zhang, Y.; Zhao, J.; Wang, Y.; Fan, Y.; Zhu, L.; Yang, Y.; Chen, X.; Lu, T.; Chen, Y.; Liu, H. Prediction of hERG K+ channel blockage using deep neural networks. Chemical biology & drug design, **2019**, *94*(5), 1973-1985.

[56]. Sun, H.; Huang, R.; Xia, M.; Shahane, S.; Southall, N.; Wang, Y. Prediction of hERG liability–using SVM classification, bootstrapping and jackknifing. Molecular informatics, **2017**, *36*(4), 1600126.

[57]. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, **2011**, *12*, 2825-2830.

[58]. Ryu, S., Lim, J., Hong, S. H., & Kim, W. Y. (2018). Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network. *arXiv*, May 28, 2018, arXiv:1805.10988v3, ver. 3. DOI:10.48550/arXiv.1805.10988

[59]. Ryu, J. Y.; Lee, M. Y.; Lee, J. H.; Lee, B. H.; Oh, K. S. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics, **2020**, *36*(10), 3049-3055.

[60]. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, **2015**, 448-456.

[61]. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: a comprehensive review. Computational Social Networks, **2019**, *6*(1), 1-23.

[62]. Lee, H.M.; Yu, M.S.; Kazmi, S.R.; Oh, S.Y.; Rhee, K.H.; Bae, M.A.; Lee, B.H.; Shin, D.S.; Oh, K.S.; Ceong, H.; Lee, D. Computational determination of hERG-related cardiotoxicity of drug candidates. BMC bioinformatics, **2019**, *20*, 67-73.

[63]. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. Match, **2006**, *56*(2), 237-248.

[64]. Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. Bioinformatics, **2019**, *35*(6), 1067-1069.

[65]. Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Research, **2021**, *49*(W1), 5-14.

[66]. O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chemistry Central Journal, **2008**, *2*(1), 1-7.

[67]. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. Journal of cheminformatics, **2011**, *3*(1), 1-14.

[68]. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum, **2013**, *8*, 31.

[69]. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, **2019**, *32*.

[70]. Harris, C.R.; Millman, K.J.; Van Der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; Kern, R. Array programming with NumPy. Nature, **2020**, *585*(7825), 357-362.

[71]. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; Van Der Walt, S.J. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, **2020**, *17*(3), 261-272.

[72]. McKinney, W.; van der Walt, S.; Millman, J. Proceedings of the 9th Python in Science Conference. **2010**

[73]. Hunter, J. D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering **2007**, *9*, 90– 95, DOI: 10.1109/MCSE.2007.55

[74]. Waskom, M. L. Seaborn: Statistical Data Visualization. Journal of Open Source Software, **2021**, *6*, 3021, DOI: 10.21105/joss.03021

[75]. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.B.; Grout, J.; Corlay, S.; Ivanov, P. Jupyter Notebooks-a publishing format for reproducible computational workflows. Elpub, **2016**, 87-90.

[76]. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. Journal of machine learning research, **2008**, *9*(11), 2579–2605.

[77]. Yu K.; Sciuto C.; Jaggi M.; Musat C.; Salzmann M. Evaluating the search phase of neural architecture search. *arXiv*, November 22, 2019, arXiv:1902.08142. ver. 3. DOI:10.48550/arXiv.1902.08142

[78]. Jiang, D.; Wu, Z.; Hsieh, C.Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. Journal of cheminformatics, **2021**, *13*(1), 1-23.