

RESEARCH

Open Access



CardioGenAI: a machine learning-based framework for re-engineering drugs for reduced hERG liability

Gregory W. Kyro^{1,2*}, Matthew T. Martin², Eric D. Watt² and Victor S. Batista^{1*}

Abstract

The link between in vitro hERG ion channel inhibition and subsequent in vivo QT interval prolongation, a critical risk factor for the development of arrhythmias such as Torsade de Pointes, is so well established that in vitro hERG activity alone is often sufficient to end the development of an otherwise promising drug candidate. It is therefore of tremendous interest to develop advanced methods for identifying hERG-active compounds in the early stages of drug development, as well as for proposing redesigned compounds with reduced hERG liability and preserved primary pharmacology. In this work, we present CardioGenAI, a machine learning-based framework for re-engineering both developmental and commercially available drugs for reduced hERG activity while preserving their pharmacological activity. The framework incorporates novel state-of-the-art discriminative models for predicting hERG channel activity, as well as activity against the voltage-gated $\text{Na}_v1.5$ and $\text{Ca}_v1.2$ channels due to their potential implications in modulating the arrhythmogenic potential induced by hERG channel blockade. We applied the complete framework to pimozide, an FDA-approved antipsychotic agent that demonstrates high affinity to the hERG channel, and generated 100 refined candidates. Remarkably, among the candidates is fluspirilene, a compound which is of the same class of drugs as pimozide (diphenylmethanes) and therefore has similar pharmacological activity, yet exhibits over 700-fold weaker binding to hERG. Furthermore, we demonstrated the framework's ability to optimize hERG, $\text{Na}_v1.5$ and $\text{Ca}_v1.2$ profiles of multiple FDA-approved compounds while maintaining the physicochemical nature of the original drugs. We envision that this method can effectively be applied to developmental compounds exhibiting hERG liabilities to provide a means of rescuing drug development programs that have stalled due to hERG-related safety concerns. Additionally, the discriminative models can also serve independently as effective components of virtual screening pipelines. We have made all of our software open-source at <https://github.com/gregory-kyro/CardioGenAI> to facilitate integration of the CardioGenAI framework for molecular hypothesis generation into drug discovery workflows.

Scientific contribution

This work introduces CardioGenAI, an open-source machine learning-based framework designed to re-engineer drugs for reduced hERG liability while preserving their pharmacological activity. The complete CardioGenAI framework can be applied to developmental compounds exhibiting hERG liabilities to provide a means of rescuing drug discovery programs facing hERG-related challenges. In addition, the framework incorporates novel state-of-the-art

*Correspondence:

Gregory W. Kyro
gregory.kyro@yale.edu
Victor S. Batista
victor.batista@yale.edu

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

discriminative models for predicting hERG, Na_v1.5 and Ca_v1.2 channel activity, which can function independently as effective components of virtual screening pipelines.

Keywords Generative AI, Deep learning, Molecular optimization, hERG, Drug discovery, Machine learning, Transformers, Graph neural networks

Introduction

There is a well-established connection between in vitro blockade of the hERG (human Ether-à-go-go-Related Gene) potassium ion channel and in vivo QT interval prolongation, where the QT interval, as recorded on electrocardiograms, indicates the time between the start of the heart's ventricular depolarization (i.e., the rapid influx of sodium ions that renders the cell's interior less negatively charge) and the end of repolarization (i.e., the restoration of the cell's membrane potential to its resting negative state) [1]. The hERG channel contributes to repolarization of the cardiac action potential by selectively allowing potassium ions to flow out of the cell following depolarization [2]. Inhibition of this channel can therefore directly disrupt cardiac repolarization, leading to prolongation of the QT interval, which consequently elevates the risk of potentially fatal arrhythmias such as Torsade de Pointes (TdP) [3]. As a result, the potential propensity of drug candidates to present hERG liabilities is subject to rigorous regulatory scrutiny, and the pharmaceutical industry devotes a significant amount of resources to identifying hERG liabilities during early, pre-clinical and clinical phases of drug development [4].

The Comprehensive In Vitro Proarrhythmia Assay (CiPA) initiative [5], supported by regulatory agencies including the U.S. Food and Drug Administration (FDA), established guidelines for evaluating the proarrhythmia risk of drugs that also incorporate the voltage-gated sodium (Na_v1.5) and calcium (Ca_v1.2) ion channels alongside the hERG channel due to observations that modulating Na_v1.5 and Ca_v1.2 channel activities may mitigate the arrhythmogenic potential induced by hERG channel blockade [6–8]. A well-known example of this phenomenon is the case of verapamil, a drug that blocks both hERG and Ca_v1.2 channels and is observed to have only a small impact on the QT interval, which is hypothesized to be due to the counteracting effects of Ca_v1.2 blockade [9]. Additionally, Ca_v1.2 blockade alone is reported to be a possible mechanism underlying undesirable blood-flow dynamics [10]. It is therefore of tremendous interest to develop highly capable methods for assessing how both prospective and currently available drugs interact with each of these three cardiac ion channels.

A multitude of experimental methods exist for in vitro determination of cardiac ion channel affinity [11–14].

However, they require synthesis of the compounds to be assayed, which is relatively time-consuming and expensive compared to in silico methods. Machine learning (ML)-based methods for predicting hERG channel activity have been extensively explored, utilizing both protein structure-based and ligand-based models [15–39]. However, structure-based predictive modeling of the hERG channel has proven to be difficult due to the channel's intricate structure, its dynamic nature encompassing multiple conformations, and the possibility of unexpected interaction sites that are not apparent in conventional structural models [40]. For these reasons, ligand-based methods currently predominate. Predictive modeling for Na_v1.5 and Ca_v1.2 channel blocking is comparatively unexplored, as the amount of available data is much less compared to that for hERG. However, recent benchmarks for predicting Na_v1.5 and Ca_v1.2 channel activity have been established [41], and increasing effort is being devoted to developing models for these channels as well [42–45].

While ML-based discriminative models for predicting hERG channel activity have tremendous potential for applications in virtual screening, extending these capabilities to molecular generation through generative artificial intelligence (AI) can overcome the constraints of the currently available molecular libraries by enabling the direct in silico development of drugs with desired activities against cardiac ion channels. Numerous generative models have already demonstrated the ability to produce molecules with prespecified drug-like properties [46–105], and there has also been work aimed at generating molecules with desired on-target potency [53, 106, 107]. Despite the progress, there has been comparatively less effort devoted to developing and applying generative models for off-target potency optimization. Moreover, the abundance of available datapoints with low hERG activity, as opposed to the general scarcity of datapoints with high on-target potency for a given target, suggests that generative models for off-target potency optimization can more effectively identify patterns in the relevant chemical space and therefore be more successful than those for on-target potency optimization, further motivating method development in this area of research.

In this work, we present an ML-based framework designed to re-engineer both developmental and commercially available drugs for reduced hERG liability while

retaining their pharmacological activity. The method utilizes a generative model to produce molecules conditioned on the molecular scaffold and physicochemical properties of the input hERG-active molecule. The generated ensemble is filtered using deep learning models for predicting hERG, Na_v1.5 and Ca_v1.2 channel activity. A chemical space representation is then constructed from the filtered generated distribution and the input molecule, where nearby molecules exhibit similar chemical properties, thus facilitating the identification of molecules with similar pharmacological activity to the input molecule but with reduced hERG channel inhibition. This approach, while not a replacement for the expertise of medicinal chemists, is highly effective at rapid molecular hypothesis generation, proposing refined candidates that can then be investigated with more expensive computational methods and experimental techniques.

Overview of CardioGenAI framework

The CardioGenAI framework combines generative and discriminative ML models to re-engineer hERG-active compounds for reduced hERG channel inhibition while preserving their pharmacological activity. A transformer decoder is trained on a dataset that we previously curated which contains approximately 5 million unique and valid SMILES strings derived from ChEMBL 33, GuacaMol

v1, MOSES, and BindingDB datasets [108–112]. The model is trained autoregressively, receiving a sequence of SMILES tokens as context as well as the corresponding molecular scaffold and physicochemical properties, and iteratively predicting each subsequent token in the sequence. Once trained, this model, which is effectively a compression of the training set, is able to generate valid molecules conditioned on a specified molecular scaffold along with a set of physicochemical properties. For an input hERG-active compound, the generation is conditioned on the scaffold and physicochemical properties of this compound (Fig. 1A). Each generated compound is subject to filtering based on activity against hERG, Na_v1.5 and Ca_v1.2 channels. Depending on the desired activity against each channel, the framework employs either classification models to include predicted non-blockers (i.e., pIC₅₀ value ≤ 5.0) or regression models to include compounds within a specified range of predicted pIC₅₀ values. Both the classification and regression models utilize the same architecture, and are trained using three feature representations of each molecule: a feature vector that is extracted from a bidirectional transformer trained on SMILES strings, a molecular fingerprint, and a graph (more details in Sect. "Data Featurization"). For each molecule in the filtered generated ensemble and the input hERG-active molecule, a feature vector is

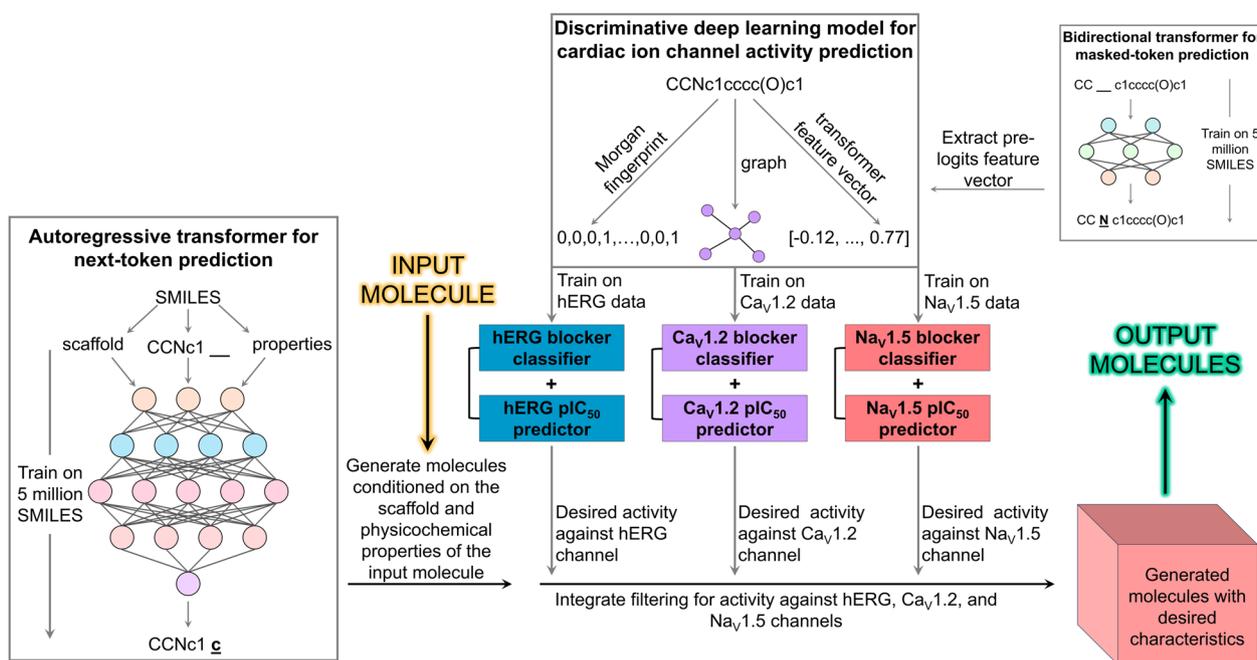


Fig. 1 The CardioGenAI framework for re-engineering hERG-active compounds. An autoregressive transformer decoder pretrained on a large dataset of SMILES strings generates compounds conditioned on the scaffold and physicochemical properties of a given input compound, and the generated ensemble is filtered based on desired activity against hERG, Na_v1.5 and Ca_v1.2 channels. Cosine similarity is calculated between a 209-dimensional descriptor vector of the input compound and that of every filtered generated compound to identify the refined candidates most chemically similar to the input compound

constructed from the 209 2D chemical descriptors available through the RDKit Descriptors module [113]. The redundant descriptors are then removed according to pairwise mutual information calculated for every possible pair of descriptors. Cosine similarity is then calculated between the processed descriptor vector of the input molecule and the descriptor vectors of every filtered generated molecule to identify the refined candidates most chemically similar to the input molecule (Fig. 1B).

Discriminative models for predicting cardiac ion channel activity

Data featurization

For training and evaluation of hERG, Na_v1.5 and Ca_v1.2 inhibition prediction models, we utilize the training and evaluation datasets included in the benchmarks recently developed by Arab et al. [41] These benchmarks are designed to assess model generalizability, enforcing a maximum fingerprint similarity cutoff between molecules in the training and evaluation sets. Multiple published models in the field have been assessed using evaluation sets that have significant overlap with the corresponding training sets [38, 114], undoubtedly yielding overoptimistic results with respect to the models' abilities to generalize. The compounds in the evaluation sets used in this work have a structural similarity, as determined by pairwise Tanimoto similarity between 2048-bit Morgan fingerprints, no greater than 0.70 to any compound in the corresponding training or validation sets. Compounds were sourced from the ChEMBL bioactivity database [115–117], PubChem [118], BindingDB [112, 119], hERGCentral [120], and the scientific literature [38, 121–123]. Each molecule is represented as a SMILES string which was canonicalized using RDKit, and labeled with the experimentally determined cardiac ion channel pIC₅₀ value. For compounds with multiple experimentally determined pIC₅₀ values, the assigned label is calculated as the mean value while retaining only those within the 95th percentile to minimize the influence of outliers. For binary classification tasks, compounds with a pIC₅₀ value greater than or equal to 5.0 are labeled as blockers. For hERG, Na_v1.5 and Ca_v1.2 channels, training sets contain 17 796 (78.3%), 1 653 (74.8%), and 641 (72.6%) datapoints, validation sets contain 4 450 (19.6%), 414 (18.7%), and 161 (18.2%) datapoints, and test sets contain 474 (2.1%), 142 (6.4%), and 81 (9.2%) datapoints, respectively. For more details regarding the curation of the datasets, we refer readers to the original paper. [41]

It is important to note that variations in experimental protocols could contribute to discrepancies in measured pIC₅₀ values for each channel due to differences in the probabilities of each channel occupying open, closed and inactivated states [124, 125]. Moreover, it has been

demonstrated that systematic differences in assay conditions, such as temperature, voltage protocols, and buffer composition, can lead to significant discrepancies in reported values. For instance, even minor deviations in experimental setup have been shown to cause variability exceeding 0.5 log units in pIC₅₀ values for the same compound across different studies [126]. Thus, given that the datasets used are curations of publicly available data that were obtained via different experimental protocols, variability in the experimental conditions and state probabilities may set an artificial limit on the predictive accuracy that models can achieve.

We found there to be a positive correlation (Pearson $r=0.256$) between hERG pIC₅₀ values and the logarithm of the partition coefficient between n-octanol and water (LogP), as well as a negative correlation (Pearson $r=-0.215$) with topological polar surface area (TPSA) (Figure S1 in Additional file 1). These findings are consistent with established medicinal chemistry knowledge that increasing polarity or reducing lipophilicity reduces hERG channel blockade [127]. Additionally, we also identified a relation between hERG pIC₅₀ values and the presence of charged nitrogen atoms within aromatic or hydrophobic groups among the molecules exhibiting the most substantial hERG activity (Figure S2 in Additional file 1).

We represent each compound as three distinct forms: a 256-dimensional feature vector that is extracted from a bidirectional transformer trained on SMILES strings, a 1024-bit Extended-Connectivity Fingerprint with a diameter of 4 bonds (ECFP4) generated using the Morgan algorithm, and a graph (Fig. 2). A bidirectional transformer is first trained for masked-token prediction on the same dataset used to train the autoregressive transformer, allowing it to develop an intricate internal representation of molecular structure and grasp the syntax of SMILES notation (more details in Sect. "Data Preparation"). After this model is fully trained, it is used as a means of extracting a context-rich feature vector as a representation of a given SMILES string. Specifically, we extract the processed vector from the penultimate layer of the model corresponding to the start token, which contains information about the entire SMILES string that contributes to the prediction of a masked token within the sequence. This information encapsulates nuanced inter-token relationships and patterns among different molecules, rendering this feature vector a powerful representation that captures important characteristics of the molecule in a high-dimensional space (more details in Sect. "Model Architectures").

In the graph representation, nodes are atoms and edges are bonds. Each node is represented as a 14-dimensional vector of atomic features: carbon indicator, nitrogen

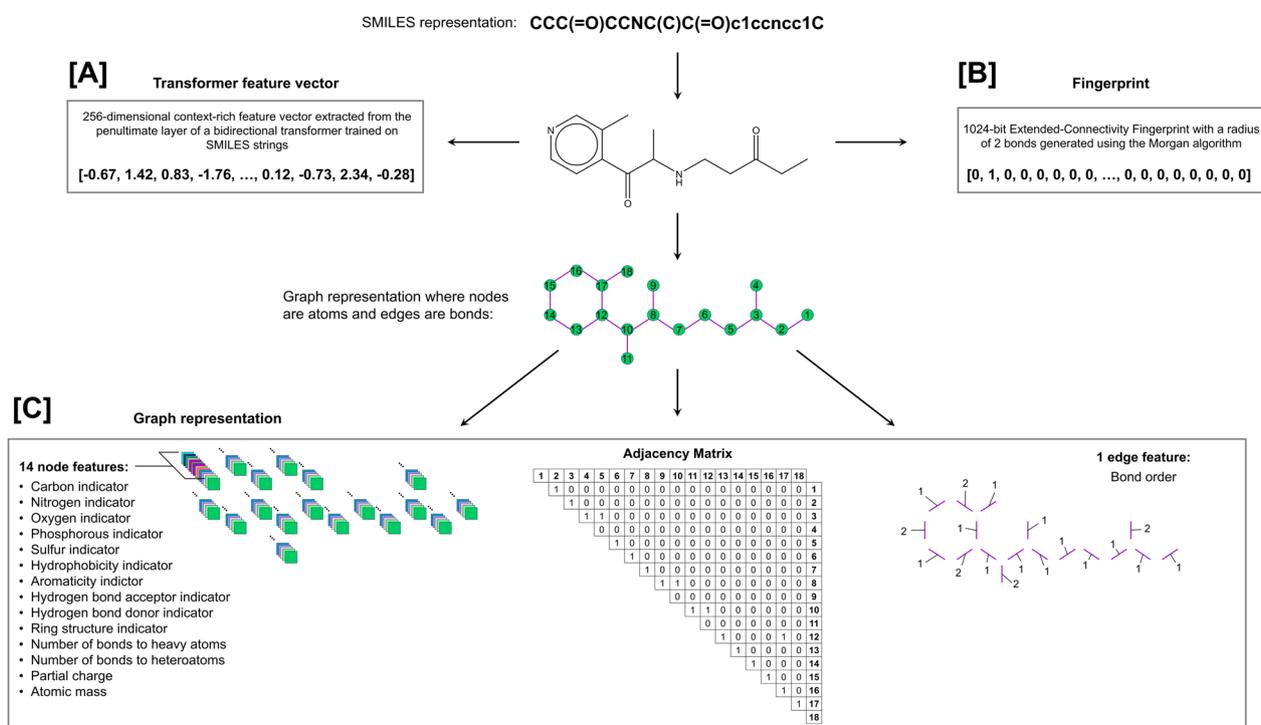


Fig. 2 Featurization of a SMILES string—CCC(=O)CCNC(C)C(=O)c1ccncc1C—for use by the CardioGenAI discriminative models. The SMILES string is represented as **A** a 256-dimensional feature vector that is extracted from the penultimate layer of a bidirectional transformer trained on SMILES strings, **B** a 1024-bit Extended-Connectivity Fingerprint with a diameter of 4 bonds (ECFP4) generated using the Morgan algorithm, and **C** a graph

indicator, oxygen indicator, phosphorous indicator, sulfur indicator, hydrophobicity indicator, aromaticity indicator, hydrogen bond acceptor indicator, hydrogen bond donor indicator, ring structure indicator, number of bonds to heavy atoms, number of bonds to heteroatoms, partial charge, and atomic mass. Each edge is labeled with the corresponding bond order.

Model Architecture

The transformer-based feature vector and the ECFP4 are each processed by separate two-layer feed-forward networks (Fig. 3B, C). For each of the two layers of the networks, the input vector undergoes a linear transformation followed by batch normalization. The normalized output is then passed through a ReLU activation function, followed by dropout with a rate of 50%.

The graph representation is processed by a graph attention network (GAT) consisting of two GAT convolutional layers (Fig. 3A). Initially, the graph is augmented with self-loops to ensure that each node's feature vector is included in its own neighborhood during feature aggregation. The first GAT layer transforms the node feature vectors through a linear operation, followed by a softmax-based attention mechanism to assign weights to the features of each node's neighbors, relative to the source

node. The output of this layer is passed through a ReLU activation function and fed to the second GAT convolutional layer which operates analogously to the first layer. After being processed by the second GAT convolutional layer, the updated node features are aggregated to form a graph-level representation using a global add pooling operation, which sums the node features across all nodes to generate a single vector that encapsulates the entire graph's information.

After each of the three input feature representations has been encoded, they are concatenated to form a combined feature vector. This combined feature vector is then passed through a two-layer feed-forward network (Fig. 3D). The first layer applies a linear transformation to the feature vector followed by batch normalization. The normalized output is then passed through a ReLU activation function followed by dropout with a rate of 50%. The output of this layer then undergoes a linear transformation to map it to the final output space.

Trainings and hyperparameters

The classification and regression models for each cardiac ion channel were trained for 200 and 100 epochs, respectively, with a batch size of 32; we trained the classification models for an additional 100 epochs because

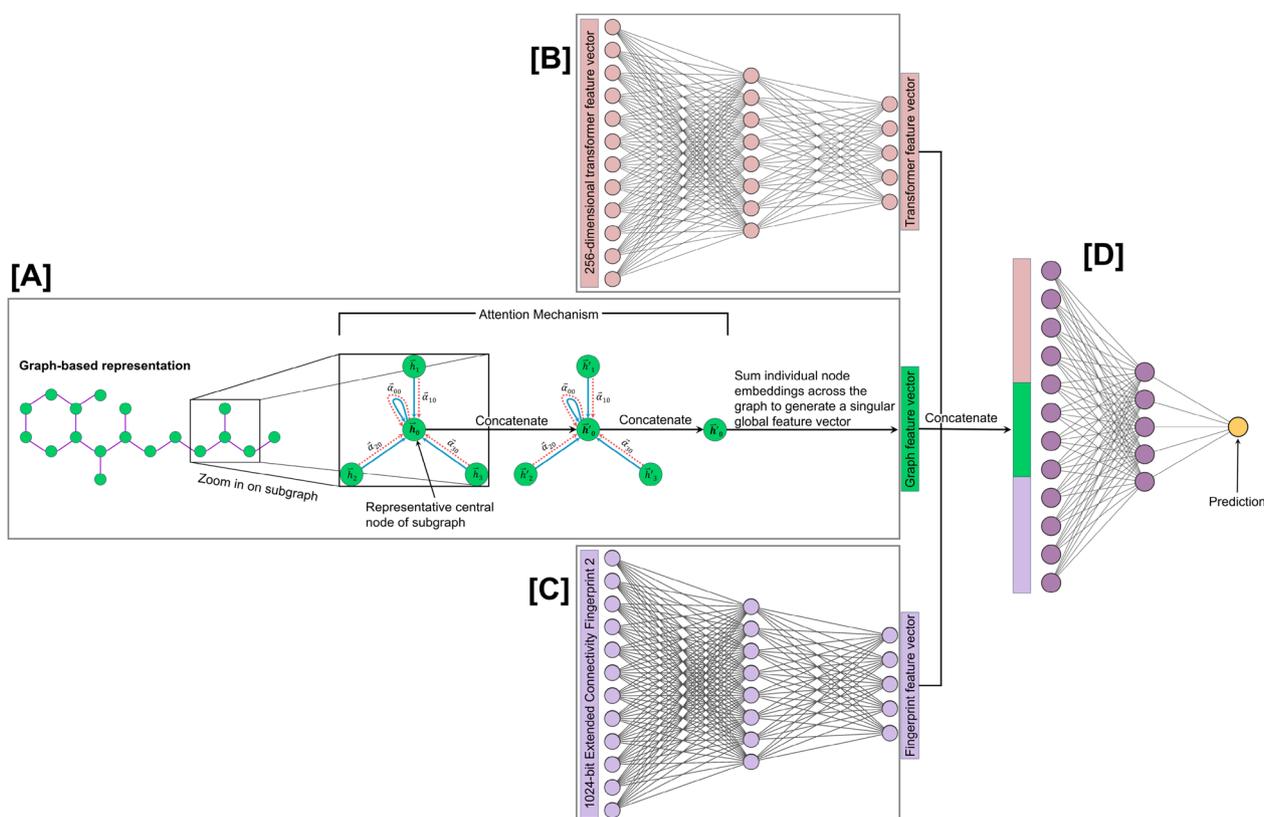


Fig. 3 Illustration of the forward pass of the CardioGenAI discriminative models. The graph representation of a given SMILES string is encoded by **A** a graph attention network (GAT). The **B** transformer-derived and **C** fingerprint feature vectors are encoded by feed-forward networks. These three encodings are then concatenated and passed to **D** a final feed-forward network to generate a prediction

the training loss had not converged after only 100 epochs (Figure S3 of Additional File 1). The AdamW optimizer, a variant of the Adam optimizer that incorporates weight decay for regularization, was used with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} to optimize the models' parameters. Additionally, L1 regularization was applied with a regularization coefficient of 1×10^{-4} to induce sparsity within the model parameters. We integrated a learning rate scheduler which monitors the validation loss and halves the learning rate if no improvement is observed for 10 consecutive epochs. To ensure stability in training and prevent gradient explosion, gradient clipping was applied with a maximum norm of 5.0. For the classification and regression models, binary cross entropy loss and mean squared error loss were used as objective functions, respectively. The model parameters used for inference are those from the epoch with the highest validation accuracy for classification and highest validation Pearson correlation for regression. Learning curves for each of the classification and regression models are reported in Figure S3 of Additional file 1.

Benchmarking against existing models

We found that utilizing all three feature representations (i.e., transformer-based feature vector, fingerprint, and graph) achieves the best performance on the hERG blocker classification benchmark compared to using any other possible combination of feature representations (Table S4 in Additional file 1), and we therefore adopt this combination of feature representations for our classification models.

We compare the performance of our classification models to the highest-performing models in the literature that have been evaluated with the benchmarks used in this work. Computed metrics include:

$$\text{Accuracy (AC)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity (SN)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity (SP)} = \frac{TN}{TN + FP} \quad (3)$$

$$F1 - \text{score (F1)} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (4)$$

$$\text{Correct Classification Rate (CCR)} = \frac{SN + SP}{2} \quad (5)$$

$$\begin{aligned} \text{Matthews Correlation Coefficient (MCC)} \\ = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \end{aligned} \quad (6)$$

where *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively. We find that our hERG blocker classification model outperforms all existing models in the literature on the hERG benchmark for binary classification (Table 1).

The improvement of our hERG blocker predictive model over previous models justifies its use within the CardioGenAI framework as opposed to other predictive models which have already been developed.

For the Na_v1.5 and Ca_v1.2 benchmarks, only the models presented by Arab et al. [41] have been evaluated, largely owing to the fact that these benchmarks have only recently been developed and the experimental data available for these channels is scarce compared to that for hERG. We find that our models demonstrate superior performance for both Na_v1.5 and Ca_v1.2 channels (Table 2). Additionally, the area under the curve (AUC) of the receiver operating characteristic for each channel is commensurate with the accuracy that our models obtain; hERG AUC is 0.88, Na_v1.5 AUC is 0.89, and Ca_v1.2 AUC is 0.95 (Figure S5B in Additional file 1).

We report the performance of our regression models in Figure S5C-E and Table S6 in Additional file 1. The Pearson correlation between true pIC₅₀ values and those predicted by our regression models are 0.67 for hERG, 0.60 for Na_v1.5, and 0.81 for Ca_v1.2 benchmarks (Figure S5C-E in Additional file 1).

In order to provide interpretability of the regression models' predictions, we calculate the correlation between predicted pIC₅₀ values and each property in a set of physicochemical properties for each of the three cardiac ion channels (Table S7 in Additional file 1). The key findings of this analysis are as follows: predicted hERG pIC₅₀ values correlate positively with the number of rotatable bonds (Pearson *r*=0.327) and LogP (*r*=0.321); predicted Na_v1.5 pIC₅₀ values correlate negatively with the number of hydrogen bond donors (*r*=−0.593) and TPSA (*r*=−0.545), while correlating positively with LogP (*r*=0.406); and predicted Ca_v1.2 pIC₅₀ values correlate positively with the number of hydrogen bond acceptors (*r*=0.621), TPSA (*r*=0.581), the number of heteroatoms

(*r*=0.555), molecular weight (*r*=0.444) and the number of rotatable bonds (*r*=0.318), while correlating negatively with the number of rings (*r*=−0.315).

Additionally, in order to ensure that the predictive abilities of our models are not artifacts of spurious correlations within the data, we perform Y-randomization tests for all discriminative models and report results in Table S8 and Figure S9 of Additional file 1.

Application to the drugcentral database of FDA-approved drugs

To demonstrate the practical utility of our classification and regression models, we applied them to the FDA-approved drugs from the DrugCentral database, offering a real-world context for assessing cardiac ion channel inhibition [130, 131]. It is important to note that many of the compounds occur in the training set of the discriminative models. Thus, predictive ability for these compounds should not be interpreted as validation of the models' predictive ability for unseen compounds. Of the 1692 unique FDA-approved drugs, we classify 504 (29.8%) to be hERG blockers (i.e., pIC₅₀ value ≥ 5.0), 764 (45.2%) to be Na_v1.5 blockers, and 400 (23.6%) to be Ca_v1.2 blockers (Figure S10A in Additional file 1). A more complete analysis of the predicted cardiac ion channel activity of the FDA-approved drugs is reported in Figure S10B of Additional file 1. In addition, we report the compounds with a predicted hERG pIC₅₀ value above 7.0 (i.e., more than 100-fold greater hERG inhibitory potency than the blocker threshold) in Table 3.

For the 11 FDA-approved compounds with a predicted hERG pIC₅₀ value greater than 7.0, the predicted pIC₅₀ values are closely aligned with those that are experimentally determined, with notable agreement in cases where the compound is not in the training set of the model (Table 3). However, for three of the compounds, namely pimoziide, astemizole, and dofetilide, each predicted hERG pIC₅₀ value differs from the corresponding experimentally determined value by about an order of magnitude. The experimentally determined pIC₅₀ values for these three compounds are among the top four highest values in the set of FDA-approved compounds, and each is greater than three standard deviations above the mean pIC₅₀ value in the training distribution. Because these high values are not well-represented in the training set, the model's tendency to regress toward the mean pIC₅₀ value likely accounts for the observed discrepancy between predicted and experimentally determined pIC₅₀ values for these three compounds (see Figure S5C in Additional File 1).

The primary mechanism of action for three of the 11 drugs is to block the hERG channel: ibutilide [134], dofetilide [135], and amiodarone [136]. Another three

of them function primarily as dopamine D2 receptor antagonists: pimozide [137], droperidol [138], and haloperidol decanoate [139]. Pimozide is reported to cause QT interval prolongation and ventricular arrhythmias due to hERG channel blockade with high specificity and affinity [140]; droperidol is reported to cause TdP due to potent hERG channel blockade [141]; haloperidol decanoate has been found to cause sudden death due to hERG channel blockade-induced QT interval prolongation. [142]

Another two of the 11 drugs function primarily as H₁-receptor antagonists: astemizole and terfenadine [143, 144]. Both of these drugs were withdrawn from the market due to hERG blockade-induced cardiac arrhythmias [145, 146]. Of the remaining three drugs of the 11, nintedanib is reported to cause side effects related to hERG channel blockade [147], halofantrine is found to cause hERG blockade-induced QT interval prolongation [148], and tolterodine is reported to cause hERG blockade-induced tachycardia and palpitations [149]. These results support the real-world application of CardioGenAI to hERG activity prediction.

Limitations of the discriminative models

While the discriminative models used in the CardioGenAI framework demonstrate robust predictive performance, certain limitations should be acknowledged. A key limitation arises from the variability in the experimental protocols used to obtain pIC₅₀ labels. These protocols often differ in assay conditions, measurement methodologies, and the probabilities of cardiac ion channels occupying open, closed, or inactivated states. Such variability introduces noise into the data and may impose an artificial upper bound on the predictive accuracy achievable by models trained on publicly available hERG data.

Additionally, the models' performance is likely influenced by the inherent biases present in the training data. For example, underrepresentation of certain chemical scaffolds or activity ranges could impact the generalizability of the models to the corresponding regions of chemical space.

Table 1 Performance of CardioGenAI for binary classification of hERG blockers compared to that of the highest-performing models in the literature on the benchmark created by Arab et al. [41]

Model	AC	SN	SP	F1	CCR	MCC
CardioGenAI	83.5	86.2	80.3	85.1	83.2	66.7
CToxPred-hERG	81.4	86.7	74.6	83.9	80.7	62.1
CardioTox	81.2	83.0	78.9	83.1	81.0	61.9
ADMETlab 2.0	71.7	71.6	71.8	73.8	71.7	43.1
ADMETsar 2.0	68.5	84.5	48.3	75.0	66.4	35.5
CardPred	56.1	52.7	60.3	57.0	56.5	13.0

^a Compounds in the test set have a structural similarity, as determined by pairwise Tanimoto similarity between 2048-bit Morgan fingerprints, no greater than 0.70 to any compound in the corresponding training or validation sets

^b The top value achieved for each metric is shown in bold

^c Accuracy (AC), sensitivity (SN), specificity (SP), F1-score (F1), correct classification rate (CCR), and Matthew's correlation coefficient (MCC) are shown

^d Results are shown for CToxPred-hERG [41], CardioTox [33], ADMETlab 2.0 [128], ADMETsar 2.0 [129], and CardPred. [21]

Table 2 Performance of CardioGenAI for binary classification of Na_v1.5 and Ca_v1.2 blockers compared to that of the models created by Arab et al. [41]

Channel	Model	AC	SN	SP	F1	CCR	MCC
Na _v 1.5	CardioGenAI	89.4	95.9	75.6	92.5	85.7	75.1
	CToxPred-Nav	81.7	85.6	73.3	86.5	79.5	58.2
Ca _v 1.2	CardioGenAI	91.4	96.2	82.8	93.5	89.5	81.0
	CToxPred-Cav	86.4	96.2	69.0	90.1	82.6	70.2

^a Compounds in the test set have a structural similarity, as determined by pairwise Tanimoto similarity between 2048-bit Morgan fingerprints, no greater than 0.70 to any compound in the corresponding training or validation sets

^b The top value achieved for each metric is shown in bold

^c Accuracy (AC), sensitivity (SN), specificity (SP), F1-score (F1), correct classification rate (CCR), and Matthew's correlation coefficient (MCC) are shown

Transformer-based models

Data preparation

The generative autoregressive transformer and the bidirectional transformer used for extracting features to be utilized by the discriminative models are both trained with a dataset that we previously curated by combining all of the unique and valid SMILES strings from ChEMBL 33, GuacaMol v1, MOSES, and BindingDB datasets [108–112]. The combined dataset initially had a vocabulary of 196 unique tokens. To reduce the size of the vocabulary and thus improve the computational efficiency of the transformer models, we removed all SMILES strings containing at least one token that appeared less than 1 000 times in the combined dataset; most of the SMILES strings that were excluded contain rare transition metals or isotopes. Of the remaining SMILES strings, the longest one contained 1 503 tokens, while 99.99% of the strings in the entire remaining dataset had 133 or fewer tokens. In order to reduce the block size of our transformer models, and thus further improve the computational efficiency,

we removed all SMILES strings from the dataset that contained more than 133 tokens. The remaining SMILES strings were then extended, if necessary, to a length of 133 using a padding token “<pad>”, and augmented with a start token “[CLS]” and an end token “[EOS]”. The processed dataset contains approximately 5.5 million SMILES strings which are randomly split into training (5 262 776 entries; 95%) and validation (276 989 entries; 5%) sets. Please refer to our previous paper for complete details regarding SMILES string preprocessing. [108]

For each SMILES string, we calculated the molecular scaffold using the Murcko algorithm [150], which identifies the core structure by removing side chains from the molecular graph, retaining the ring systems and the linkers connecting them. We also calculated ten physicochemical properties for each SMILES string: molecular weight, number of rings, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, TPSA, number of heteroatoms, LogP, number of stereocenters, and formal charge.

Table 3 Analysis of the FDA-approved compounds from the DrugCentral database with a predicted hERG pIC₅₀ value above 7.0

Drug Name	Pharmacological Indication	Mechanism of Action	FDA Approval Status	Predicted hERG pIC ₅₀	In Training Set	Experimentally Determined hERG pIC ₅₀
Nintedanib	Idiopathic pulmonary fibrosis	Kinase inhibitor	Approved	8.234	yes	8.585
Ibutilide	Atrial fibrillation, atrial flutter	hERG channel blocker	Approved	7.977	yes	8.000
Pimozide	Tourette's disorder	Dopamine D2 receptor antagonist	Approved	7.629	yes	8.520
Halofantrine	Severe malaria	Forms toxic complexes with ferritoporphyrin IX	Approved	7.588	no	7.398
Astemizole	Allergy symptoms	H ₁ -receptor antagonist	Withdrawn due to concerns of arrhythmias	7.562	yes	8.538
Tolterodine	Overactive bladder	Muscarinic receptor antagonist	Approved	7.311	no	7.886
Droperidol	Nausea and vomiting in surgical and diagnostic procedures	Dopamine D2 receptor antagonist	Approved	7.300	yes	7.495
Dofetilide	Atrial fibrillation, atrial flutter	hERG channel blocker	Approved	7.164	yes	8.194
Haloperidol decanoate	Schizophrenia, psychotic disorders, Tourette's disorder	Dopamine D2 receptor antagonist	Approved	7.149	no	6.921
Amiodarone	Recurrent ventricular fibrillation, recurrent hemodynamically unstable ventricular tachycardia	hERG channel blocker	Approved	7.127	yes	7.523
Terfenadine	Allergic rhinitis, hay fever, allergic skin disorders	H ₁ -receptor antagonist	Withdrawn due to concerns of arrhythmias	7.022	yes	7.252

^a Information regarding the pharmacological indication, mechanism of action, and FDA approval status for each drug is obtained from DrugBank [132, 133]

^b Experimentally determined hERG pIC₅₀ values are obtained from the hERG dataset curated by Arab et al. [41]

Model architectures

For a given SMILES string, the autoregressive transformer considers the sequence of the SMILES string, the molecular scaffold, and the set of physicochemical properties, while the bidirectional transformer only considers the sequence. For both models, tokens in the sequence are embedded using a learnable embedding table, where each token in the vocabulary corresponds to a learnable embedding vector. The positions of the tokens in the sequence are embedded using a separate learnable embedding table, where each index in the sequence corresponds to a learnable embedding vector that allows the model to account for a given token's position in the sequence and capture sequential context within the SMILES string. For the autoregressive transformer, the set of physicochemical properties is mapped to the embedding dimension via a learnable linear transformation, and the molecular scaffold is embedded using a learnable embedding table analogous to that used for the token embeddings. For both models, all embeddings, each with an embedding dimension of 256, are summed to create a combined feature representation, and then dropout is applied with a rate of 10%.

The transformer architecture used consists of eight sequential blocks, each beginning with layer normalization to stabilize the input. This is followed by a self-attention mechanism, where query (Q), key (K), and value (V) vectors are computed for each input token, attention scores are derived via a scaled dot product of Q and K vectors, and the softmax function normalizes these scores to obtain weights that modulate the aggregation of V , effectively capturing the magnitude with which each token will attend to every other token in the sequence. The self-attention mechanism is executed multiple times in parallel through what is referred to as multi-head attention. The models used in this work employ eight attention heads, where each head uses its own set of learned linear transformations to generate Q , K , and V vectors for each token in the sequence, allowing the model to simultaneously focus on different aspects of the input across the various heads. Representative attention maps for the autoregressive and bidirectional transformers are reported in Figures S11 and S12 of Additional file 1.

The outputs of all attention heads are concatenated and passed through a learned linear transformation to generate the final output of the multi-head attention mechanism. A residual connection then merges this output with the initial block input. The resulting data tensor then undergoes another layer normalization and progresses through a two-layer feed-forward network with a 10% dropout rate and GeLU activation,

before reintegration with its pre-normalized state. The final step involves another layer normalization, followed by a linear transformation that projects the data tensor onto the vocabulary space, generating a logits vector (i.e., the unnormalized log probabilities for each token in the vocabulary). When using the trained bidirectional transformer to derive feature vectors to be utilized by the discriminative models, the data tensor is extracted immediately prior to the final linear transformation, and the vector corresponding to the start token is used as the feature vector.

Trainings and hyperparameters

The autoregressive transformer is trained for next-token prediction, and the bidirectional transformer is trained for masked-token prediction where each token in a given SMILES sequence has a 15% probability of being selected; of these, 80% are replaced with a mask token "<MASK>"; 10% are replaced with a random token from the vocabulary, and the remaining 10% are left unchanged. Both models were trained for 100 epochs with a batch size of 512. The Sophia optimizer was used with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-1} , [151] and cross entropy loss was used as the objective function for both models. The model parameters used for inference are those from the last epoch of training. Learning curves for the autoregressive and bidirectional transformers are reported in Figure S13 of Additional file 1.

Molecular generation

The autoregressive transformer is used to generate SMILES strings, conditioned on both a molecular scaffold and a set of ten physicochemical properties. To rigorously evaluate the model's ability to generate molecules with prespecified physicochemical properties, we fix one property at a time to a discrete value while the other nine properties are sampled using a random uniform distribution within ranges of acceptable values based on ADMETlab 2.0 guidelines for medicinal chemistry [128]. This procedure is performed for 500 molecules per fixed property value. For example, we generate 500 molecules conditioned on a molecular weight of $400 \frac{\text{g}}{\text{mol}}$ and another 500 conditioned on a molecular weight of $600 \frac{\text{g}}{\text{mol}}$ to assess the model's ability to generate molecules with a targeted molecular weight. We repeat this approach for each physicochemical property, and observe that the model is able to successfully generate molecular distributions that satisfy the prespecified criteria (Figure S14A-I in Additional file 1). We also demonstrate the model's ability to generate molecules conditioned on multiple discrete physicochemical property values simultaneously (e.g., TPSA of 50 \AA^2 [2] and molecular weight of $350 \frac{\text{g}}{\text{mol}}$), validating its utility and justifying its use within

the CardioGenAI framework (Figure S14J in Additional file 1).

Complete CardioGenAI framework

High-level description of the workflow

The fundamental objective of the CardioGenAI framework is to re-engineer hERG-active compounds for reduced hERG activity while preserving their pharmacological action. Within the framework, the autoregressive transformer first generates valid molecules conditioned on the molecular scaffold and physicochemical properties of the input hERG-active molecule, which are filtered based on desired activity against hERG, Na_v1.5 and Ca_v1.2 channels using the discriminative models. The input molecule and each filtered generated molecule are then converted into 209-dimensional chemical descriptor vectors which are refined by removing the redundant descriptors according to pairwise mutual information between every possible descriptor pair. Cosine similarity is then calculated between the descriptor vector of the input molecule and the descriptor vectors of every filtered generated molecule to identify the molecules most chemically similar to the input molecule but with desired activity against each of the cardiac ion channels.

Case study: optimizing the FDA-approved drug pimozide for reduced hERG activity

Pimozide is an FDA-approved antipsychotic agent that is used to treat Tourette's syndrome as well as various other psychiatric disorders [152]. Its main pharmacodynamic action is to blockade dopamine D2 receptors on neurons in the central nervous system (CNS); it also has various effects on other CNS receptor systems which are not fully characterized [137]. There are many reports linking the use of pimozide to QT interval prolongation and ventricular arrhythmias [153, 154], and there are multiple reported instances of sudden, unexpected deaths of patients receiving pimozide [155].

It was initially observed clinically that only a very low dose of pimozide is necessary to produce QT interval prolongation, suggesting that it binds to one or more cardiac potassium ion channels with high affinity [153]. Subsequent experimental validation indicated pimozide's high affinity to the hERG channel, evidenced by its potent inhibitory effect with an IC₅₀ value of approximately 18 nM [140].

Because of pimozide's proarrhythmic effects, it is contraindicated in patients with congenital long QT syndrome, patients with a history of cardiac arrhythmias, patients taking other drugs that prolong the QT interval, and patients with known hypokalemia (i.e., low potassium levels) or hypomagnesemia (i.e., low magnesium levels) [155]. It is therefore of tremendous interest to

develop safer alternatives to pimozide that minimize its hERG activity while retaining its therapeutic efficacy.

In this work, we apply the CardioGenAI framework to re-engineer pimozide for reduced hERG inhibition while preserving its pharmacological activity. The experimentally determined pIC₅₀ value of pimozide for the hERG channel is 8.520, and the value that our regression model predicts is 7.629, a difference (0.891 pIC₅₀) which is sufficiently small to be attributable to variance in experimental protocols used to obtain labels [156]. Our objective is to generate compounds with similar pharmacological properties, but with predicted hERG channel pIC₅₀ values less than 6.0.

We therefore condition the molecular generation on the scaffold and physicochemical properties of pimozide, and filter out molecules with a predicted hERG channel pIC₅₀ value greater than or equal to 6.0. This procedure is performed until 100 compounds are generated, which takes approximately one minute using an NVIDIA GeForce RTX 4050 GPU. We then compute descriptor vectors for pimozide and the filtered generated molecules, and then calculate the cosine similarity between the descriptor vector of pimozide and those of the generated molecules. In practice, many more molecules can be generated to create a molecular library for further screening.

We calculate the ten previously described physicochemical properties for pimozide, the 100 filtered generated molecules, and the molecules in the transformer training set, and then perform principal component analysis (PCA) to construct a lower-dimensional chemical space in which we can visually compare the filtered generated molecules to pimozide in relation to the broader transformer training set. Plotting the first two PCs reveals that the filtered generated molecules are closely aligned to pimozide, indicating that our framework successfully navigates the initially vast chemical space to propose compounds with similar physicochemical characteristics to pimozide but with reduced hERG activity (Fig. 4A; Figure S15 in Additional file 1). Additionally, the distribution of predicted pIC₅₀ values of the generated compounds ranges from 4.64 to 6.00 with a mean value of 5.59, indicating significant reductions in hERG activity (Fig. 4B). The most similar generated molecules to pimozide are reported in Table S16 of Additional file 1.

We analyze each of the 100 generated refined compound with respect to all of the compounds provided in the DrugCentral Postgres v14.5 database to identify any compounds approved by either the FDA, the European Medicines Agency (EMA), or the Pharmaceuticals and Medical Devices Agency of Japan (PMDA) [130, 131]. Remarkably, among the 100 filtered generated compounds is fluspirilene, a compound that belongs to the

same class of drugs as pimozone (diphenylmethanes) and therefore presents a highly similar pharmacological profile [157]. Moreover, the experimental hERG pIC_{50} value of fluspirilene is 5.638 (predicted: 5.785), as compared to 8.520 (predicted: 7.629) for pimozone (Fig. 5), indicating a reduction in hERG activity by over 700-fold.

The reduced hERG activity of fluspirilene compared to pimozone can be attributed to the presence of an aromatic nitrogen-containing heterocyclic group in pimozone, which is absent in fluspirilene (Fig. 5). Aromaticity increases the basicity of the nitrogen, allowing for protonation and stronger electrostatic and π -cation interactions with the hERG channel. This aligns with prior literature and our observations (Sect. "Data Featurization") that basic, aromatic nitrogens are significant contributors to hERG activity [127].

This case study demonstrates the ability of the CardioGenAI framework to re-engineer a hERG-active compound for reduced hERG activity while preserving its pharmacological activity.

Additional applications of the complete framework for hERG activity optimization

In addition to re-engineering pimozone, we also apply the CardioGenAI framework to nintedanib, ibutilide, halofantrine, and astemizole. Collectively, including pimozone, these five compounds are those among the set of FDA-approved compounds provided by DrugCentral that have the highest predicted pIC_{50} values against the hERG channel. We show that for each drug, the framework is able to successfully generate compounds with similar physicochemical profiles and with significantly reduced activity against the hERG channel (Fig. 6).

Applications of the complete framework for $Na_v1.5$ and $Ca_v1.2$ activity optimization

Moreover, given that modulating $Na_v1.5$ and $Ca_v1.2$ channel activities may mitigate the arrhythmogenic potential induced by hERG channel blockade [6–8], and considering that activity against each of these two channels alone can present problems related to the cardiac action potential [10, 45], we demonstrate the ability of the framework to optimize compounds for enhanced $Na_v1.5$ and $Ca_v1.2$ profiles. Specifically, we assess the capabilities of the framework with respect to four independent objectives: (1) Increase the $Na_v1.5$ activity of a compound that has high hERG activity but low $Na_v1.5$ activity; (2) Increase the $Ca_v1.2$ activity of a compound that has high hERG activity but low $Ca_v1.2$ activity; (3) Decrease the $Na_v1.5$ activity of a compound that has high $Na_v1.5$ activity; (4) Decrease the $Ca_v1.2$ activity of a compound that has high $Ca_v1.2$ activity. For cases (1) and (2), we chose to re-engineer ibutilide, which has a

predicted pIC_{50} for hERG, $Na_v1.5$, and $Ca_v1.2$ of 7.98, 4.24 and 4.02, respectively. For case (3), we chose venetoclax, which has a predicted $Na_v1.5$ pIC_{50} of 6.72. For case (4), we chose itraconazole, which inhibits $Ca_v1.2$ with a predicted pIC_{50} of 9.17. The CardioGenAI framework is able to successfully improve the cardiac ion channel activity by at least one order of magnitude in each case for every generated refined compound while ensuring that the generated compounds are physicochemically similar to the respective input drug. The results for each of these four cases are presented in Fig. 7.

Customizing the CardioGenAI framework for company-specific industrial applications

Pharmaceutical companies have begun to leverage generative AI-based methods for specific tasks within the earlier stages of drug discovery pipelines [158]. In order to facilitate integration of CardioGenAI into drug discovery workflows, all of the software is entirely open-source and the framework is designed to be easily customizable. Companies can therefore incorporate desired functionality, and retrain all of the models on their internal data. It is expected that large pharmaceutical companies will significantly benefit from retraining the models, given that their internal data is likely more comprehensive and subject to significantly less experimental variance than the publicly available datasets used to initially train the models.

With respect to the incorporation of additional functionality into the framework, CardioGenAI is designed such that predictive models can easily be integrated into the filtering phase along with the cardiac ion channel activity prediction models. For instance, a team of medicinal chemists will likely adhere to synthesis-related criteria; a rule-based filter, or a model fit to these criteria, can easily be incorporated. The objective of such a model could be to identify compounds that can be produced given an initial compound and feasible synthetic pathways, or to predict a synthetic accessibility score for a given compound. In theory, any predictive model can be integrated into the framework (e.g., for predicting on-target activity, solubility, metabolic stability, bioavailability, etc.).

Because synthesizability is arguably the most important characteristic of a proposed compound, additional steps can be taken, aside from incorporating more models, to ensure that the proposed compounds are in accordance with a company's specific synthesis capabilities. For instance, the dataset used to train the generative autoregressive transformer could be curated to contain only compounds that a company deems sufficiently synthesizable, thereby biasing the generative component of the framework to only propose compounds that are akin to

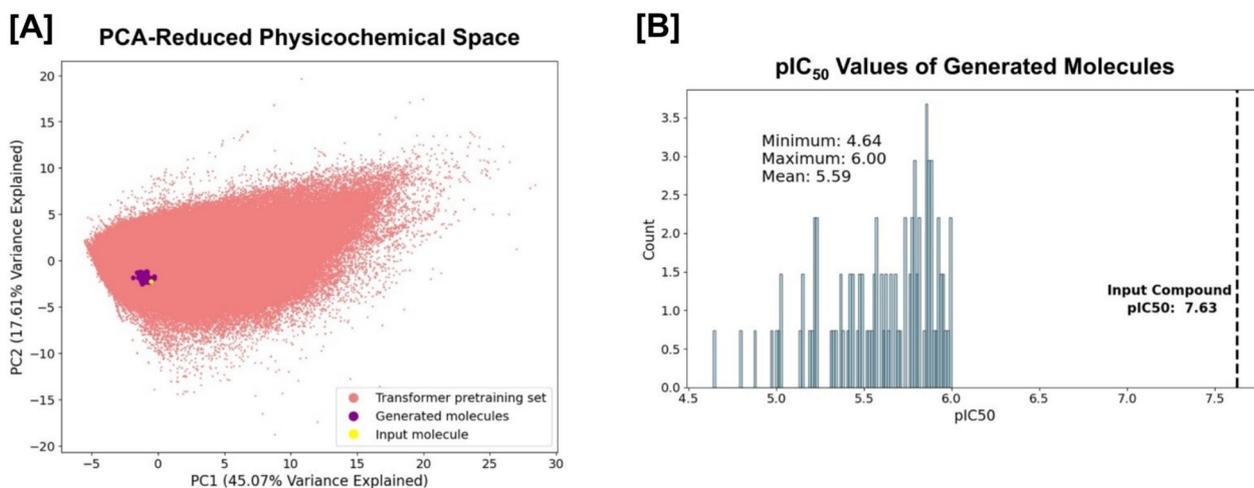


Fig. 4 Visualization of the CardioGenAI framework applied to pimozide. The input molecule (pimozide), the 100 generated refined molecules, and the molecules in the training set for the transformer-based models (approximately 5 million datapoints), are projected into a principal component analysis (PCA)-reduced physicochemical-based space, shown in **(A)**. Pimozide is colored yellow, the generated refined compounds are colored purple, and the compounds in the training set of the transformer-based models are colored red. The first two principal components explain 45.07% and 17.61% of the total variance, respectively. Clearly, the CardioGenAI framework is able to identify the region of physicochemical space corresponding to compounds that are similar to pimozide, yet exhibit significantly reduced activity against the hERG channel. The density of predicted pIC_{50} values against the hERG channel of the generated refined compounds as compared to that of pimozide is shown in **(B)**. The distribution of generated compounds exhibits a maximum predicted pIC_{50} value of 6.00, with a mean of 5.59 and minimum of 4.64

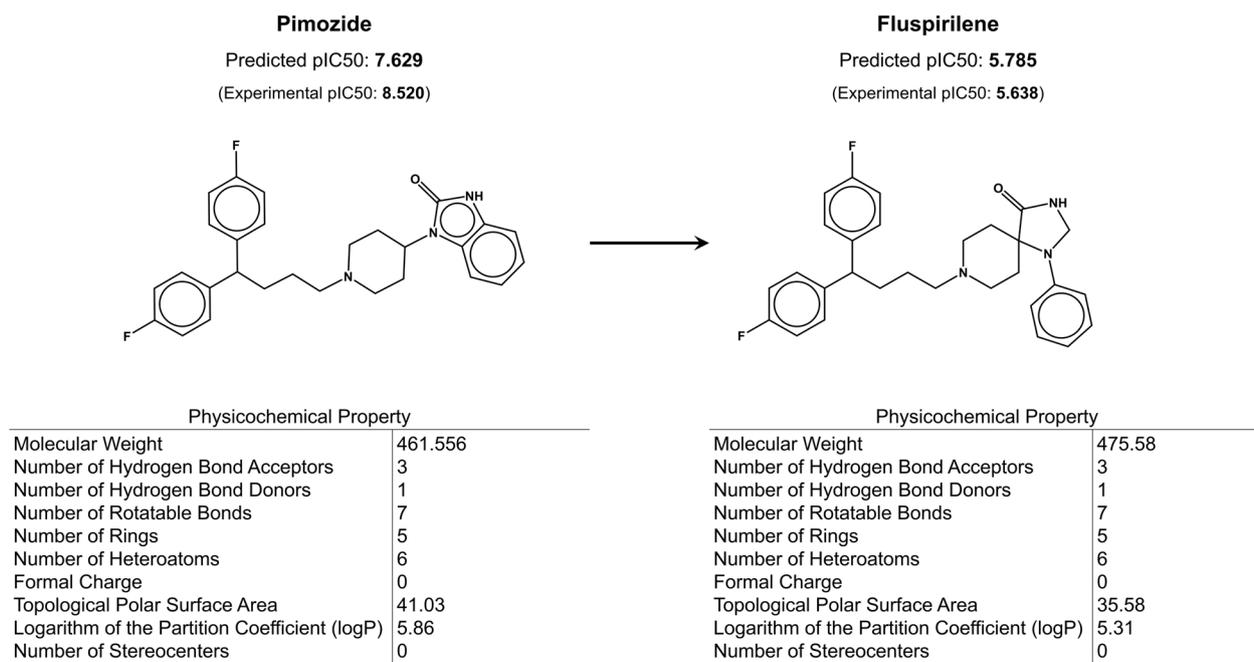


Fig. 5 CardioGenAI framework applied to pimozide, an FDA-approved antipsychotic drug that has an experimental hERG pIC_{50} value of 8.520 (predicted: 7.629), and is reported to cause hERG channel blockade-induced QT interval prolongation and arrhythmias. CardioGenAI proposes 100 molecules, and among them is fluspirilene, a compound that belongs to the same class of drugs as pimozide but exhibits over 700-fold weaker binding to hERG (experimental pIC_{50} value is 5.638)

those that satisfy these synthesizability standards. Additionally, rather than defining the chemical space based on RDKit descriptors to identify molecules that are physicochemically similar to the input molecule, the space can be designed such that nearby molecules are easily synthesizable.

In the current implementation, RDKit is used to validate the proposed molecules generated by the framework, ensuring that molecular representations conform to basic valence and bonding rules. However, it does not assess chemical plausibility beyond these criteria. As such, some structures may be valid according to RDKit but exhibit features that are chemically improbable. To address this, the framework can easily be augmented with additional criteria applied at the generation stage

to enforce properties such as thermodynamic stability or broader chemical plausibility. These enhancements allow users to refine the generative process further, ensuring that proposed compounds align with expectations.

Summary

Although numerous generative models have demonstrated the ability to produce molecules with prespecified drug-like properties, as well as molecules with desired on-target potency, there has been comparatively less effort devoted to developing and applying generative models for off-target potency optimization. In this work, we present an ML-based framework for re-engineering hERG-active compounds for reduced hERG activity while preserving their pharmacological activity.

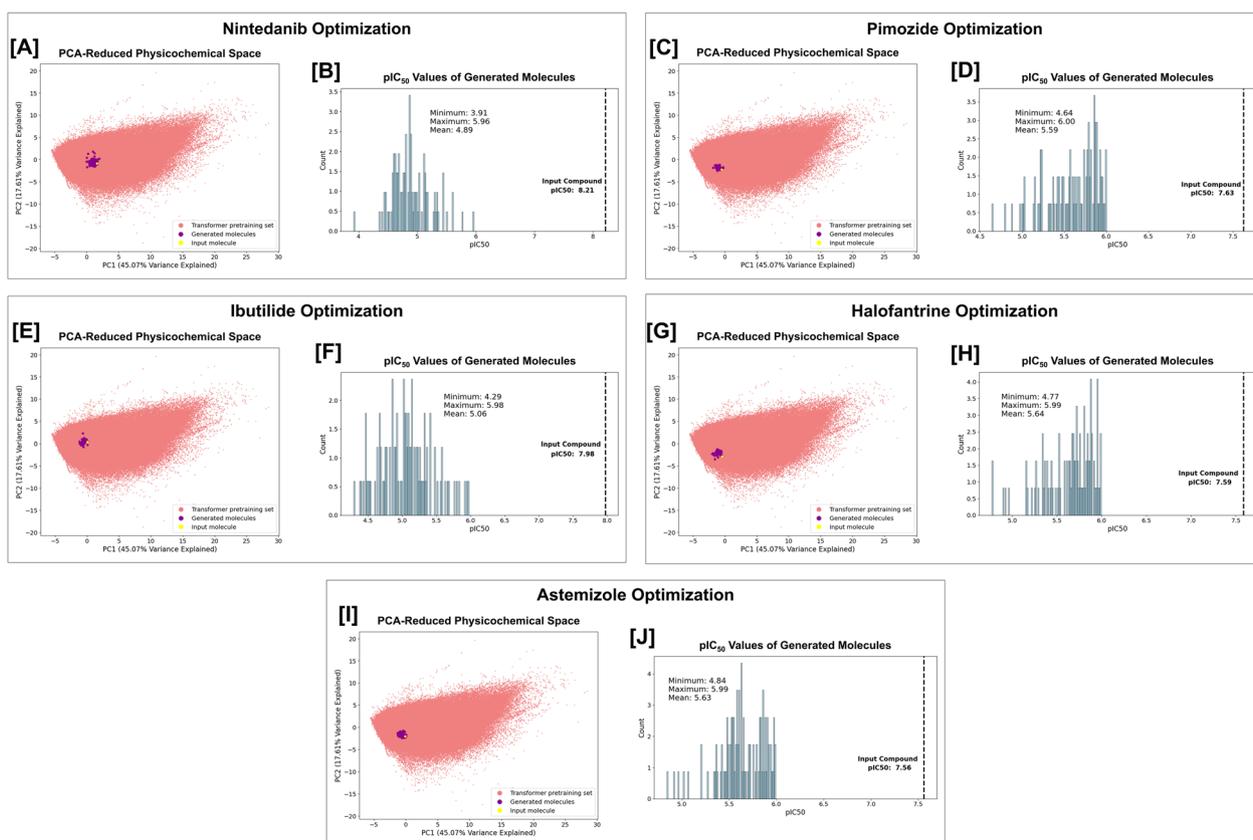


Fig. 6 Visualization of the CardioGenAI framework applied to nintedanib (A, B), pimoziide (C, D), ibutilide (E, F), halofantrine (G, H), and astemizole (I, J). In each application, the specified maximum predicted hERG pIC₅₀ value of any of the generated compounds was set to 6.00. For each optimization, the input molecule, the 100 generated refined molecules, and the molecules in the training set for the transformer-based models (approximately 5 million datapoints), are projected into a principal component analysis (PCA)-reduced physicochemical-based space. The input compound is colored yellow, the generated refined compounds are colored purple, and the compounds in the training set of the transformer-based models are colored red. The first two principal components explain 45.07% and 17.61% of the total variance, respectively. In each case, the CardioGenAI framework is able to identify the region of physicochemical space corresponding to compounds that are similar to the input compound, yet exhibit significantly reduced activity against the hERG channel. The densities of predicted pIC₅₀ values against the hERG channel of the generated refined compounds as compared to that of the respective input compound are shown in [B]. Relevant metrics are shown on each plot

The method utilizes an autoregressive transformer-based generative model to produce molecules conditioned on the molecular scaffold and set of physicochemical properties of the input molecule. The generated ensemble is filtered based on hERG, Na_v1.5 and Ca_v1.2 activity using state-of-the-art discriminative deep learning models. A physicochemical-based space is then constructed from the filtered generated distribution and the input molecule, where nearby molecules have similar physicochemical profiles, thus facilitating the identification of molecules with similar pharmacological activity to the input molecule but with reduced hERG liability. We applied the framework to pimozide, an FDA-approved antipsychotic agent that demonstrates high affinity to the hERG channel, and generated a compound of the same class of drugs that has a significantly lower hERG pIC₅₀ value as indicated by both predicted and experimental values. Furthermore, we demonstrated the framework's ability to optimize hERG, Na_v1.5 and Ca_v1.2 profiles of multiple FDA-approved compounds while maintaining the physicochemical nature of the original drugs. In

addition, the state-of-the-art performances of the hERG, Na_v1.5, and Ca_v1.2 activity prediction models support their independent utility as effective components of virtual screening campaigns.

Technical implementation details

The transformer-based models and the feed-forward networks in the discriminative models were built using PyTorch [159]. The parameters of the transformer-based models were optimized using the Sophia optimizer [151]. The GAT components of the discriminative models were built using PyTorch Geometric [160]. The hyperparameters of the discriminative models were optimized using Optuna [161]. The hyperparameters that were optimized include: batch size, learning rate, weight decay, the number of GAT attention heads used in the graph model, the output dimension of the GAT mechanism used in the graph model, and the dropout rate applied to the fully connected components of the complete architecture. SMILES canonicalization, as well as the calculations of physicochemical properties and molecular scaffolds were

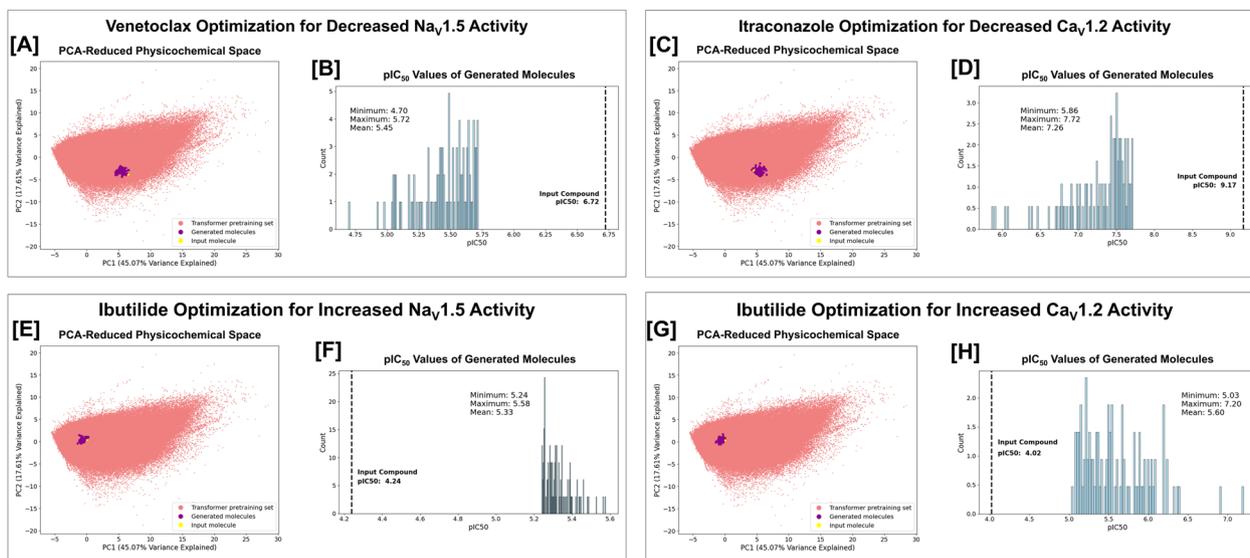


Fig. 7 Visualization of the CardioGenAI framework applied to venetocloax (A, B), itraconazole (C, D), and ibutilide (E–H). In each case, the specified predicted cardiac ion channel pIC₅₀ value for each of the generated compounds is set to be at least an improvement of one order of magnitude compared to that of the input compound. For each optimization, the input molecule, the 100 generated refined molecules, and the molecules in the training set for the transformer-based models (approximately 5 million datapoints), are projected into a principal component analysis (PCA)-reduced physicochemical-based space. The input compound is colored yellow, the generated refined compounds are colored purple, and the compounds in the training set of the transformer-based models are colored red. The first two principal components explain 45.07% and 17.61% of the total variance, respectively. For venetocloax, which has a predicted Na_v1.5 pIC₅₀ of 6.72, we reduce the Na_v1.5 pIC₅₀ by at least one order of magnitude for each generated compound (B). For itraconazole, which inhibits Ca_v1.2 with a predicted pIC₅₀ of 8.72, we reduce the Ca_v1.2 pIC₅₀ by at least one order of magnitude for each generated compound (D). For ibutilide, which has a predicted pIC₅₀ for hERG, Na_v1.5, and Ca_v1.2 of 7.98, 4.24 and 4.02, respectively, we independently increase the Na_v1.5 pIC₅₀ by at least one order of magnitude for each generated compound (F) and increase the Ca_v1.2 pIC₅₀ by at least one order of magnitude for each generated compound (H). In each case, the CardioGenAI framework is able to identify the region of physicochemical space corresponding to compounds that are similar to the input compound, yet exhibit significantly improved activity against the respective cardiac ion channel. The densities of predicted pIC₅₀ values of the generated refined compounds against the respective cardiac ion channel are shown. Relevant metrics are shown on each plot

performed using RDKit [113]. Scikit-learn was used to calculate pairwise mutual information between chemical features and cosine similarity between descriptor vectors, as well as to perform PCA [162].

Abbreviations

hERG	Human Ether-à-go-go-Related Gene
TdP	Torsade de Pointes
CiPA	The Comprehensive In Vitro Proarrhythmia Assay
FDA	U.S. Food and Drug Administration
Na _v 1.5	Voltage-gated sodium ion channel subtype 1.5
Ca _v 1.2	Voltage-gated calcium ion channel subtype 1.2
ML	Machine learning
AI	Artificial intelligence
LogP	Logarithm of the partition coefficient between n-octanol and water
TPSA	Topological polar surface area
ECFP4	Extended-Connectivity Fingerprint with a diameter of 4 bonds
GAT	Graph attention network
AC	Accuracy
SN	Sensitivity
SP	Specificity
CCR	Correct classification rate
MCC	Matthew's correlation coefficient
AUC	Area under the curve
Q	Query vector
K	Key vector
V	Value vector
CNS	Central nervous system
PCA	Principal component analysis

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-025-00976-8>.

Supplementary Material 1. Details regarding the datasets used, model trainings, additional analyses of the models, and the refined drug candidates.

Acknowledgements

We acknowledge financial support from the National Science Foundation Graduate Research Fellowship under Grant DGE-2139841 [GWK], from the National Science Foundation Engines Development Award: Advancing Quantum Technologies (CT) under Award Number 2302908 [VSB], and from the CCI Phase I: National Science Foundation Center for Quantum Dynamics on Modular Quantum Devices (CQD-MQD) under Award Number 2124511 [VSB]. Additionally, we acknowledge seed funding from Yale University. We also acknowledge high-performance computer time from the National Energy Research Scientific Computing Center and from the Yale University Faculty of Arts and Sciences High Performance Computing Center. We also thank Todd A. Wisialowski, Peter J. Kilfoil, and Nathaniel Woody for their valuable comments and expert insights regarding the manuscript.

Author contributions

G.W.K., M.T.M., E.D.W., V.S.B. conceived the idea; G.W.K., M.T.M., E.D.W. designed research; G.W.K. developed software; G.W.K. performed research; G.W.K., M.T.M., E.D.W. analyzed data; G.W.K., M.T.M., E.D.W. wrote the paper; V.S.B. provided feedback on the paper. All authors have given approval to the final version of the manuscript.

Funding

National Science Foundation Graduate Research Fellowship: Grant DGE-2139841. National Science Foundation Engines Development Award – Advancing Quantum Technologies (CT): Award Number 2302908. CCI Phase I – National Science Foundation Center for Quantum Dynamics on Modular Quantum Devices (CQD-MQD): Award Number 2124511.

Availability of data and materials

All of our software is available as open-source at <https://github.com/gregory-kyro/CardioGenAI>. Users can easily run the complete CardioGenAI framework, perform inference with the discriminative models, and reproduce the figures in this manuscript. Additionally, we provide all of the data we use, as well as the parameters for each of our trained models.

Declarations

Competing interests

The authors declare no competing interests.

Author details

¹Department of Chemistry, Yale University, New Haven, CT 06511, USA. ²Drug Safety Research & Development, Pfizer Research & Development, Groton, CT 06340, USA.

Received: 11 August 2024 Accepted: 21 February 2025

Published online: 05 March 2025

References

- Food; Administration, D.; Health, U. D. o.; Services, H. Guidance for industry. E14 clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs. 2005. <http://www.fda.gov/cder/guidance/6922fnl.pdf>.
- Jones DK, Liu F, Vaidyanathan R, Eckhardt LL, Trudeau MC, Robertson GA (2014) hERG 1b is critical for human cardiac repolarization. *Proc Natl Acad Sci* 111(50):18073–18077. <https://doi.org/10.1073/pnas.1414945111>
- Sanguinetti MC, Tristani-Firouzi M (2006) hERG potassium channels and cardiac arrhythmia. *Nature* 440(7083):463–469. <https://doi.org/10.1038/nature04710>
- Sun D, Gao W, Hu H, Zhou S (2022) Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B* 12(7):3049–3062. <https://doi.org/10.1016/j.apsb.2022.02.002>
- Sager PT, Gintant G, Turner JR, Pettit S, Stockbridge N (2014) Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium. *Am Heart J* 167(3):292–300. <https://doi.org/10.1016/j.ahj.2013.11.004>
- Kowalska M, Nowaczyk J, Nowaczyk A (2020) K(V)11.1, Na(V)1.5, and Ca(V)1.2 transporter proteins as antitarget for drug cardiotoxicity. *Int J Mol Sci*. <https://doi.org/10.3390/ijms21218099>
- Warner B, Hoffmann P (2002) Investigation of the potential of clozapine to cause torsade de pointes. *Adverse Drug React Toxicol Rev* 21:189–203
- Bril A, Gout B, Bonhomme M, Landais L, Faivre J-F, Linee P, Poyser RH, Ruffolo R (1996) Combined potassium and calcium channel blocking activities as a basis for antiarrhythmic efficacy with low proarrhythmic risk: experimental profile of BRL-32872. *J Pharmacol Exp Ther* 276(2):637–646
- Britton OJ, Abi-Gerges N, Page G, Ghetti A, Miller PE, Rodriguez B (2017) Quantitative comparison of effects of dofetilide, sotalolol, quinidine, and verapamil between human ex vivo trabeculae and in silico ventricular models incorporating inter-individual action potential variability. *Front Physiol* 8:597. <https://doi.org/10.3389/fphys.2017.00597>FromNLM
- Balasubramanian B, Imredy JP, Kim D, Penniman J, Lagrutta A, Salata JJ (2009) Optimization of Cav1.2 screening with an automated planar patch clamp platform. *J Pharmacol Toxicol Methods* 59(2):62–72. <https://doi.org/10.1016/j.vascn.2009.02.002>
- Meyer T, Boven K-H, Günther E, Fejtł M (2004) Micro-electrode arrays in cardiac safety pharmacology: a novel tool to study QT interval prolongation. *Drug Saf* 27:763–772
- Finlayson K, Turnbull L, January CT, Sharkey J, Kelly JS (2001) [3H] dofetilide binding to hERG transfected membranes: a potential high throughput preclinical screen. *Eur J Pharmacol* 430(1):147–148
- Dorn A, Hermann F, Ebnet A, Bothmann H, Trube G, Christensen K, Apfel C (2005) Evaluation of a high-throughput fluorescence assay

- method for hERG potassium channel inhibition. *J Biomol Screen* 10(4):339–347
14. Cheng CS, Alderman D, Kwash J, Dessaint J, Patel R, Lescoe MK, Kinrade MB, Yu W (2002) A high-throughput hERG potassium channel function assay: an old assay with a new look. *Drug Dev Ind Pharm* 28(2):177–191
 15. Creanza TM, Delre P, Ancona N, Lentini G, Saviano M, Mangiardi GF (2021) Structure-based prediction of hERG-related cardiotoxicity: a benchmark study. *J Chem Inf Model* 61(9):4758–4770. <https://doi.org/10.1021/acs.jcim.1c00744>
 16. Kalyanamoorthy S, Lamothe SM, Hou X, Moon TC, Kurata HT, Houghton M, Barakat KH (2020) A structure-based computational workflow to predict liability and binding modes of small molecules to hERG. *Sci Rep* 10(1):16262. <https://doi.org/10.1038/s41598-020-72889-5>
 17. Krishna S, Borrel A, Huang R, Zhao J, Xia M, Kleinstreuer N (2022) High-throughput chemical screening and structure-based models to predict hERG inhibition. *Biology* 11(2):209
 18. Hari Narayana Moorthy NS, Karthikeyan C, Manivannan E (2021) Multi-algorithm based machine learning and structural pattern studies for hERG ion channel blockers mediated cardiotoxicity prediction. *Chemom Intell Lab Syst* 208:104213. <https://doi.org/10.1016/j.chemo.2020.104213>
 19. Ryu JY, Lee MY, Lee JH, Lee BH, Oh K-S (2020) DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics* 36(10):3049–3055. <https://doi.org/10.1093/bioinformatics/btaa075> (accessed 2/3/2024)
 20. Kim H, Nam H (2020) hERG-Att: Self-attention-based deep neural network for predicting hERG blockers. *Comput Biol Chem* 87:107286. <https://doi.org/10.1016/j.compbiolchem.2020.107286>
 21. Lee H-M, Yu M-S, Kazmi SR, Oh SY, Rhee K-H, Bae M-A, Lee BH, Shin D-S, Oh K-S, Ceong H et al (2019) Computational determination of hERG-related cardiotoxicity of drug candidates. *BMC Bioinformatics* 20(10):250. <https://doi.org/10.1186/s12859-019-2814-5>
 22. Zhang Y, Zhao J, Wang Y, Fan Y, Zhu L, Yang Y, Chen X, Lu T, Chen Y, Liu H (2019) Prediction of hERG K⁺ channel blockage using deep neural networks. *Chem Biol Drug Des* 94(5):1973–1985. <https://doi.org/10.1111/cbdd.13600>
 23. Choi K-E, Balupuri A, Kang NS (2020) The study on the hERG blocker prediction using chemical fingerprint analysis. *Molecules* 25(11):2615
 24. Sramshetty VB, Nguyen D-T, Martinez NJ, Southall NT, Simeonov A, Zakharov AV (2020) Critical assessment of artificial intelligence methods for prediction of hERG channel inhibition in the “big data” era. *J Chem Inf Model* 60(12):6007–6019. <https://doi.org/10.1021/acs.jcim.0c00884>
 25. Meng J, Zhang L, Wang L, Li S, Xie D, Zhang Y, Liu H (2021) TSSF-hERG: a machine-learning-based hERG potassium channel-specific scoring function for chemical cardiotoxicity prediction. *Toxicology* 464:153018. <https://doi.org/10.1016/j.tox.2021.153018>
 26. Ogura K, Sato T, Yuki H, Honma T (2019) Support Vector Machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II. *Sci Rep* 9(1):12220. <https://doi.org/10.1038/s41598-019-47536-3>
 27. Liu M, Zhang L, Li S, Yang T, Liu L, Zhao J, Liu H (2020) Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. *Toxicol Lett* 332:88–96. <https://doi.org/10.1016/j.toxlet.2020.07.003>
 28. Hu J, Huang M, Ono N, Chen-Izu Y, Izu LT, Kanaya S (2019) Cardiotoxicity prediction based on integrated hERG database with molecular convolution model. *IEEE Int Conf Bioinform Biomed*. <https://doi.org/10.1109/BIBM47256.2019.8983163>
 29. Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, Fang J, Cheng F (2019) Deep learning-based prediction of drug-induced cardiotoxicity. *J Chem Inf Model* 59(3):1073–1084. <https://doi.org/10.1021/acs.jcim.8b00769>
 30. Wang T, Sun J, Zhao Q (2023) Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. *Comput Biol Med* 153:106464. <https://doi.org/10.1016/j.compbiomed.2022.106464>
 31. Zhang X, Mao J, Wei M, Qi Y, Zhang JZH (2022) HergSPred: accurate classification of hERG blockers/nonblockers with machine-learning models. *J Chem Inf Model* 62(8):1830–1839. <https://doi.org/10.1021/acs.jcim.2c00256>
 32. Kim H, Park M, Lee I, Nam H (2022) BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbac211>
 33. Karim A, Lee M, Balle T, Sattar A (2021) CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics* 13(1):60. <https://doi.org/10.1186/s13321-021-00541-z>
 34. Chen Y, Yu X, Li W, Tang Y, Liu G (2023) In silico prediction of hERG blockers using machine learning and deep learning approaches. *J Appl Toxicol* 43(10):1462–1475. <https://doi.org/10.1002/jat.4477>
 35. Shan M, Jiang C, Chen J, Qin L-P, Qin J-J, Cheng G (2022) Predicting hERG channel blockers with directed message passing neural networks. *RSC Adv* 12(6):3423–3430. <https://doi.org/10.1039/D1RA07956E>
 36. Delre P, Lavado GJ, Lamanna G, Saviano M, Roncaglioni A, Benfenati E, Mangiardi GF, Gadaleta D (2022) Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. *Front Pharmacol*. <https://doi.org/10.3389/fphar.2022.951083>
 37. Ding W, Nan Y, Wu J, Han C, Xin X, Li S, Liu H, Zhang L (2022) Combining multi-dimensional molecular fingerprints to predict the hERG cardiotoxicity of compounds. *Comput Biol Med* 144:105390. <https://doi.org/10.1016/j.compbiomed.2022.105390>
 38. Konda LSK, Keerthi Praba S, Kristam R (2019) hERG liability classification models using machine learning techniques. *Comput Toxicol* 12:100089. <https://doi.org/10.1016/j.comtox.2019.100089>
 39. Feng H, Wei G-W (2023) Virtual screening of DrugBank database for hERG blockers using topological Laplacian-assisted AI models. *Comput Biol Med* 153:106491. <https://doi.org/10.1016/j.compbiomed.2022.106491>
 40. Butler A, Helliwell MV, Zhang Y, Hancox JC, Dempsey CE (2020) An update on the structure of hERG. *Front Pharmacol*. <https://doi.org/10.3389/fphar.2019.01572>
 41. Arab I, Egghe K, Laukens K, Chen K, Barakat K, Bittremieux W (2023) Benchmarking of small molecule feature representations for hERG, Nav1.5, and Cav1.2 cardiotoxicity prediction. *J Chem Info Model*. <https://doi.org/10.1021/acs.jcim.3c01301>
 42. Kong W, Huang W, Peng C, Zhang B, Duan G, Ma W, Huang Z (2023) Multiple machine learning methods aided virtual screening of Nav1.5 inhibitors. *J Cell Mol Med* 27(2):266–276. <https://doi.org/10.1111/jcmm.17652>
 43. Arab I, Barakat K, ToxTree: descriptor-based machine learning models for both hERG and Nav1.5 cardiotoxicity liability predictions. 2021; p [arXiv:2112.13467](https://arxiv.org/abs/2112.13467).
 44. Chen L, Jiang J, Dou B, Feng H, Liu J, Zhu Y, Zhang B, Zhou T, Wei G-W (2023) Machine learning study of the extended drug–target interaction network informed by pain related voltage-gated sodium channels. *Pain*. <https://doi.org/10.1097/j.pain.0000000000003089>
 45. Llanos MA, Enrique N, Esteban-López V, Scioli-Montoto S, Sánchez-Benito D, Ruiz ME, Milesi V, López DE, Talevi A, Martín P, Gavernet L (2023) A combined ligand- and structure-based virtual screening to identify novel Nav1.2 blockers in vitro patch clamp validation and in vivo anti-convulsant activity. *J Chem Info Model* 63(22):7083–7096. <https://doi.org/10.1021/acs.jcim.3c00645>
 46. Segler MH, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci* 4(1):120–131
 47. Urbina F, Lowden CT, Culbertson JC, Ekins S (2022) MegaSyn: integrating generative molecular design, automated analog designer, and synthetic viability prediction. *ACS Omega* 7(22):18699–18713
 48. Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G (2018) Generative recurrent networks for de novo drug design. *Mol Inf* 37(1–2):1700111
 49. Xu M, Ran T, De CH (2021) novo molecule design through the molecular generative model conditioned by 3D information of protein binding sites. *J Chem Inf Model* 61(7):3240–3254
 50. Arús-Pous J, Blaschke T, Ulander S, Reymond J-L, Chen H, Engkvist O (2019) Exploring the GDB-13 chemical space using deep generative models. *J Cheminform* 11(1):1–14
 51. Yonchev D, Bajorath J (2020) DeepCOMO: from structure–activity relationship diagnostics to generative molecular design using the

- compound optimization monitor methodology. *J Comput Aided Mol Des* 34:1207–1218
52. Grisoni F, Moret M, Lingwood R, Schneider G (2020) Bidirectional molecule generation with recurrent neural networks. *J Chem Inf Model* 60(3):1175–1183
 53. Zhang J, De CH (2022) novo molecule design using molecular generative models constrained by ligand–protein interactions. *J Chem Inf Model* 62(14):3291–3306
 54. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. *J Cheminform* 11(1):1–13
 55. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G (2020) Generative molecular design in low data regimes. *Nat Mach Intell* 2(3):171–180
 56. Li X, Xu Y, Yao H, Lin K (2020) Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. *J Cheminform* 12(1):1–13
 57. Merk D, Friedrich L, Grisoni F, De SG (2018) novo design of bioactive small molecules by artificial intelligence. *Mol Inf* 37(1–2):1700153
 58. Tan X, Jiang X, He Y, Zhong F, Li X, Xiong Z, Li Z, Liu X, Cui C, Zhao Q (2020) Automated design and optimization of multitarget schizophrenia drug candidates by deep learning. *Eur J Med Chem* 204:112572
 59. Bjerrum EJ, Threlfall R. Molecular generation with recurrent neural networks (RNNs). 2017. arXiv preprint arXiv:1705.04612
 60. Kotsias P-C, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ (2020) Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat Mach Intell* 2(5):254–265
 61. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminform* 9(1):1–14
 62. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. *Sci Adv* 4(7):eaap7885
 63. Blaschke T, Engkvist O, Bajorath J, Chen H (2020) Memory-assisted reinforcement learning for diverse molecular de novo design. *J Cheminform* 12(1):1–17
 64. Yoshimori A, Kawasaki E, Kanai C, Tasaka T (2020) Strategies for design of molecular structures with a desired pharmacophore using deep reinforcement learning. *Chem Pharm Bull* 68(3):227–233
 65. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A (2020) REINVENT 2.0: an AI tool for de novo drug design. *J Chem Inf Model* 60(12):5918–5922
 66. Korshunova M, Huang N, Capuzzi S, Radchenko DS, Savych O, Moroz YS, Wells CI, Willson TM, Tropsha A, Isayev O (2022) Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Commun Chem* 5(1):129
 67. Popova M, Shvets M, Oliva J, Isayev O. MolecularRNN: generating realistic molecular graphs with optimized properties. 2019. arXiv preprint arXiv:1905.13372.
 68. Bian Y, Wang J, Jun JJ, Xie X-Q (2019) Deep convolutional generative adversarial network (dcGAN) models for screening and design of small molecules targeting cannabinoid receptors. *Mol Pharm* 16(11):4451–4460
 69. Méndez-Lucio O, Baillif B, Clevert D-A, Rouquié D, De WJ (2020) novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat Commun* 11(1):10
 70. De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs. 2018. arXiv preprint arXiv:1805.11973
 71. Tsujimoto Y, Hiwa S, Nakamura Y, Oe Y, Hiroyasu T. L-MolGAN: An improved implicit generative model for large molecular graphs. 2021.
 72. Wang J, Chu Y, Mao J, Jeon H-N, Jin H, Zeb A, Jang Y, Cho K-H, Song T, NoDe KT (2022) novo molecular design with deep molecular generative models for PPI inhibitors. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbac285>
 73. Song T, Ren Y, Wang S, Han P, Wang L, Li X, Rodríguez-Patón A (2023) DNMG: deep molecular generative model by fusion of 3D information for de novo drug design. *Methods* 211:10–22. <https://doi.org/10.1016/j.ymeth.2023.02.001>
 74. Bai Q, Tan S, Xu T, Liu H, Huang J, Yao X (2020) MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbaa161>
 75. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A (2018) Reinforced adversarial neural computer for de novo molecular design. *J Chem Inf Model* 58(6):1194–1204. <https://doi.org/10.1021/acs.jcim.7b00690>
 76. Lee YJ, Kahng H, Kim SB (2021) Generative adversarial networks for de novo molecular design. *Mol Inf* 40(10):2100045. <https://doi.org/10.1002/minf.202100045>
 77. Putin E, Asadulaev A, Vanhaelen Q, Ivanenkov Y, Aladinskaya AV, Aliper A, Zhavoronkov A (2018) Adversarial threshold neural computer for molecular de novo design. *Mol Pharm* 15(10):4386–4397. <https://doi.org/10.1021/acs.molpharmaceut.7b01137>
 78. Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G (2019) From target to drug: generative modeling for the multimodal structure-based ligand design. *Mol Pharm* 16(10):4282–4291. <https://doi.org/10.1021/acs.molpharmaceut.9b00634>
 79. Prykhodko O, Johansson SV, Kotsias P-C, Arús-Pous J, Bjerrum EJ, Engkvist O, Chen H (2019) A de novo molecular generation method using latent vector based generative adversarial network. *J Cheminform* 11(1):74. <https://doi.org/10.1186/s13321-019-0397-9>
 80. Abbasi M, Santos BP, Pereira TC, Sofia R, Monteiro NRC, Simões CJV, Brito RMM, Ribeiro B, Oliveira JL, Arrais JP (2022) Designing optimized drug candidates with generative adversarial network. *J Cheminform* 14(1):40. <https://doi.org/10.1186/s13321-022-00623-6>
 81. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci* 4(2):268–276. <https://doi.org/10.1021/acscentsci.7b00572>
 82. Lim J, Ryu S, Kim JW, Kim WY (2018) Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform* 10(1):31. <https://doi.org/10.1186/s13321-018-0286-7>
 83. Wang S, Song T, Zhang S, Jiang M, Wei Z, Li Z (2022) Molecular substructure tree generative model for de novo drug design. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbab592>
 84. Kang S, Cho K (2019) Conditional molecular design with deep generative models. *J Chem Inf Model* 59(1):43–52. <https://doi.org/10.1021/acs.jcim.8b00263>
 85. Lim J, Hwang S-Y, Moon S, Kim S, Kim WY (2020) Scaffold-based molecular design with a graph generative model. *Chem Sci* 11(4):1153–1164. <https://doi.org/10.1039/C9SC04503A>
 86. Dollar O, Joshi N, Beck DAC, Pfaendtner J (2021) Attention-based generative models for de novo molecular design. *Chem Sci* 12(24):8362–8372. <https://doi.org/10.1039/D1SC01050F>
 87. Krishnan SR, Bung N, Vangala SR, Srinivasan R, Bulusu G, De RA (2022) Novo structure-based drug design using deep learning. *J Chem Inf Model* 62(21):5100–5109. <https://doi.org/10.1021/acs.jcim.1c01319>
 88. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A et al (2019) Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* 37(9):1038–1040. <https://doi.org/10.1038/s41587-019-0224-x>
 89. Nesterov VI, Wieser M, Roth V. 3DMolNet: a generative network for molecular structures. ArXiv 2020, abs/2010.06477.
 90. Skalic M, Jiménez J, Sabbadin D, De Fabritiis G (2019) Shape-based generative modeling for de novo drug design. *J Chem Inf Model* 59(3):1205–1214. <https://doi.org/10.1021/acs.jcim.8b00706>
 91. Hong SH, Ryu S, Lim J, Kim WY (2020) Molecular generative model based on an adversarially regularized autoencoder. *J Chem Inf Model* 60(1):29–36. <https://doi.org/10.1021/acs.jcim.9b00694>
 92. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A (2017) The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8(7):10883–10890. <https://doi.org/10.18632/oncotarget.14073>
 93. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm* 14(9):3098–3104. <https://doi.org/10.1021/acs.molpharmaceut.7b00346>
 94. Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, Bozdaganyan M, Aliper A, Zhavoronkov A, Kadurin A (2018) Entangled conditional adversarial autoencoder for de novo drug

- discovery. *Mol Pharm* 15(10):4398–4405. <https://doi.org/10.1021/acs.molpharmaceut.8b00839>
95. Winter R, Montanari F, Steffen A, Briem H, Noé F, Clevert D-A (2019) Efficient multi-objective molecular optimization in a continuous latent space. *Chem Sci* 10(34):8016–8024. <https://doi.org/10.1039/C9SC01928F>
96. Gao K, Nguyen DD, Tu M, Wei G-W (2020) Generative network complex for the automated generation of drug-like molecules. *J Chem Inf Model* 60(12):5682–5698. <https://doi.org/10.1021/acs.jcim.0c00599>
97. Sattarov B, Baskin II, Horvath D, Marcou G, Bjerrum EJ, De VA (2019) Novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *J Chem Inf Model* 59(3):1182–1196. <https://doi.org/10.1021/acs.jcim.8b00751>
98. Mao J, Wang J, Zeb A, Cho K-H, Jin H, Kim J, Lee O, Wang Y, No KT (2023) Transformer-based molecular generative model for antiviral drug design. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.3c00536>
99. Wei L, Fu N, Song Y, Wang Q, Hu J (2023) Probabilistic generative transformer language models for generative design of molecules. *J Cheminform* 15(1):88. <https://doi.org/10.1186/s13321-023-00759-z>
100. Wang J, Mao J, Wang M, Le X, Wang Y (2023) Explore drug-like space with deep generative models. *Methods* 210:52–59. <https://doi.org/10.1016/j.jymeth.2023.01.004>
101. Grechishnikova D (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci Rep* 11(1):321. <https://doi.org/10.1038/s41598-020-79682-4>
102. Kim H, Na J, Lee WB (2021) Generative chemical transformer: neural machine learning of molecular geometric structures from chemical language via attention. *J Chem Inf Model* 61(12):5804–5814. <https://doi.org/10.1021/acs.jcim.1c01289>
103. Wang W, Wang Y, Zhao H, Sciabola S. A Transformer-based generative model for de novo molecular design. 2022; p [arXiv:2210.08749](https://arxiv.org/abs/2210.08749).
104. Chen Y, Wang Z, Wang L, Wang J, Li P, Cao D, Zeng X, Ye X, Sakurai T (2023) Deep generative model for drug design from protein target sequence. *J Cheminform* 15(1):38. <https://doi.org/10.1186/s13321-023-00702-2>
105. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD (2022) MolGPT: molecular generation using a transformer-decoder model. *J Chem Inf Model* 62(9):2064–2076. <https://doi.org/10.1021/acs.jcim.1c00600>
106. Pang C, Qiao J, Zeng X, Zou Q, Wei L (2023) Deep generative models in de novo drug molecule generation. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.3c01496>
107. Guan J, Qian WW, Peng X, Su Y, Peng J, Ma J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. 2023. [arXiv preprint arXiv:2303.03543](https://arxiv.org/abs/2303.03543)
108. Kyro GW, Morgunov A, Brent RI, Batista VS (2024) ChemSpaceAL: an efficient active learning methodology applied to protein-specific molecular generation. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.3c01456>
109. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 47(D1):D930–D940
110. Brown N, Fiscato M, Segler MH, Vaucher AC (2019) GuacaMol: benchmarking models for de novo molecular design. *J Chem Inf Model* 59(3):1096–1108
111. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M (2020) Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front Pharmacol* 11:565644
112. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2006) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 35(suppl 1):D198–D201. <https://doi.org/10.1093/nar/gkl999>
113. Landrum G. Rdkit: Open-source cheminformatics software. 2016.
114. Liu L-L, Lu J, Lu Y, Zheng M-Y, Luo X-M, Zhu W-L, Jiang H-L, Chen K-X (2014) Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. *Acta Pharmacol Sin* 35(8):1093–1102. <https://doi.org/10.1038/aps.2014.35>
115. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40(D1):D1100–D1107. <https://doi.org/10.1093/nar/gkr777>
116. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S et al (2013) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42(D1):D1083–D1090. <https://doi.org/10.1093/nar/gkt1031>
117. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E et al (2016) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954. <https://doi.org/10.1093/nar/gkw1074>
118. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2020) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 49(D1):D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>
119. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2015) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44(D1):D1045–D1053. <https://doi.org/10.1093/nar/gkv1072>
120. hERGCentral: a large database to store, retrieve, and analyze compound-human ether-à-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *ASSAY Drug Dev Technol* 2011;9(6):580–588. <https://doi.org/10.1089/adt.2011.0425>
121. Didziapetris R, Lanevskij K (2016) Compilation and physicochemical classification analysis of a diverse hERG inhibition database. *J Comput Aided Mol Des* 30(12):1175–1188. <https://doi.org/10.1007/s10822-016-9986-0>
122. Doddareddy MR, Klaasse EC, Ijzerman AP, Bender A (2010) Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *ChemMedChem* 5(5):716–729. <https://doi.org/10.1002/cmdc.201000024>
123. Munawar S, Vandenberg JI, Jabeen I (2019) Molecular docking guided grid-independent descriptor analysis to probe the impact of water molecules on conformational changes of hERG inhibitors in drug trapping phenomenon. *Int J Mol Sci* 20(14):3385
124. Gomis-Tena J, Brown BM, Cano J, Trenor B, Yang PC, Saiz J, Clancy CE, Romero L (2020) When does the IC(50) accurately assess the blocking potency of a drug? *J Chem Inf Model* 60(3):1779–1790. <https://doi.org/10.1021/acs.jcim.9b01085FromNLM>
125. Escobar F, Gomis-Tena J, Saiz J, Romero L (2022) Automatic modeling of dynamic drug-hERG channel interactions using three voltage protocols and machine learning techniques: a simulation study. *Comput Methods Programs Biomed* 226:107148. <https://doi.org/10.1016/j.cmpb.2022.107148>
126. Elkins RC, Davies MR, Brough SJ, Gavaghan DJ, Cui Y, Abi-Gerges N, Mirams GR (2013) Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment. *J Pharmacol Toxicol Methods* 68(1):112–122. <https://doi.org/10.1016/j.jvasc.2013.04.007>
127. Jamieson C, Moir EM, Rankovic Z, Wishart G (2006) Medicinal chemistry of hERG optimizations: highlights and hang-ups. *J Med Chem* 49(17):5029–5046. <https://doi.org/10.1021/jm060379l>
128. Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, Yin M, Zeng X, Wu C, Lu A et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res* 49(W1):W5–W14. <https://doi.org/10.1093/nar/gkab255>
129. Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, Li W, Liu G, Tang Y (2018) admetsAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* 35(6):1067–1069. <https://doi.org/10.1093/bioinformatics/bty707>
130. Avram S, Bologna CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, Curpan R, Halip L, Bora A, Yang JJ et al (2021) DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res* 49(D1):D1160–d1169. <https://doi.org/10.1093/nar/gkaa997>
131. Ursu O, Holmes J, Knockel J, Bologna CG, Yang JJ, Mathias SL, Nelson SJ, Oprea TI (2016) DrugCentral: online drug compendium. *Nucleic Acids Res* 45(D1):D932–D939. <https://doi.org/10.1093/nar/gkw993>
132. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34:D668–672. <https://doi.org/10.1093/nar/gkj067>

133. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36:D901–906. <https://doi.org/10.1093/nar/gkm958>
134. Murray KT (1998) Ibutilide. *Circulation* 97(5):493–497
135. Mounsey JP, DiMarco JP (2000) Dofetilide. *Circulation* 102(21):2665–2670
136. Mason JW (1987) Amiodarone. *N Engl J Med* 316(8):455–466
137. Finder R, Brogden R, Sawyer PR, Speight T, Spencer R, Avery G (1976) Pimozide: a review of its pharmacological properties and therapeutic uses in psychiatry. *Drugs* 12:1–40
138. Henzi I, Sonderegger J, Tramer MR (2000) Efficacy, dose-response, and adverse effects of droperidol for prevention of postoperative nausea and vomiting. *Can J Anesth* 47:537–551
139. Beresford R, Ward A (1987) Haloperidol decanoate: a preliminary review of its pharmacodynamic and pharmacokinetic properties and therapeutic use in psychosis. *Drugs* 33:31–49
140. Kang J, Wang L, Cai F, Rampe D (2000) High affinity blockade of the HERG cardiac K⁺ channel by the neuroleptic pimozide. *Eur J Pharmacol* 392(3):137–140. [https://doi.org/10.1016/S0014-2999\(00\)00123-0](https://doi.org/10.1016/S0014-2999(00)00123-0)
141. Drolet B, Zhang S, Deschênes D, Rail J, Nadeau S, Zhou Z, January CT, Turgeon J (1999) Droperidol lengthens cardiac repolarization due to block of the rapid component of the delayed rectifier potassium current. *J Cardiovasc Electrophysiol* 10(12):1597–1604. <https://doi.org/10.1111/j.1540-8167.1999.tb00224.x>
142. Lin Y, Sun I-W, Liu S-I, Chen C-Y, Hsu C-C (2009) QTc prolongation during concurrent treatment with depot antipsychotics and high-dose amisulpride: a report of 2 cases. *J Intern Med* 20(6):544–549
143. Richards D, Brogden R, Heel R, Speight T, Avery G (1984) Astemizole: a review of its pharmacodynamic properties and therapeutic efficacy. *Drugs* 28:38–61
144. Badwan AA, Al Kaysi HN, Owais LB, Salem MS, Arafat TA. Terfenadine. In: *Analytical Profiles of Drug Substances*, Vol. 19; Elsevier, 1990; pp 627–662.
145. Zhou Z, Vorperian VR, Gong Q, Zhang S, January CT (1999) Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole. *J Cardiovasc Electrophysiol* 10(6):836–843. <https://doi.org/10.1111/j.1540-8167.1999.tb00264.x>
146. Suessbrich H, Waldegger S, Lang F, Busch A (1996) Blockade of HERG channels expressed in *Xenopus* oocytes by the histamine receptor antagonists terfenadine and astemizole. *FEBS Lett* 385(1–2):77–80
147. Huang Z, Li H, Zhang Q, Lu F, Hong M, Zhang Z, Guo X, Zhu Y, Li S, Liu H (2017) Discovery of indolinone-based multikinase inhibitors as potential therapeutics for idiopathic pulmonary fibrosis. *ACS Med Chem Lett* 8(11):1142–1147. <https://doi.org/10.1021/acsmedchemlett.7b00164>
148. Traebert M, Dumotier B, Meister L, Hoffmann P, Dominguez-Estevéz M, Suter W (2004) Inhibition of hERG K⁺ currents by antimalarial drugs in stably transfected HEK293 cells. *Eur J Pharmacol* 484(1):41–48. <https://doi.org/10.1016/j.ejphar.2003.11.003>
149. Wang N, Yang Y, Wen J, Fan X-R, Li J, Xiong B, Zhang J, Zeng B, Shen J-W, Chen G-L (2022) Molecular determinants for the high-affinity blockade of human ether-à-go-go-related gene K⁺ channel by tolterodine. *J Cardiovasc Pharmacol* 80(5):679–689. <https://doi.org/10.1097/fjc.0000000000001336>
150. Bemis GW, Murcko MA (1996) The properties of known drugs 1 Molecular frameworks. *J Med Chem* 39(15):2887–2893. <https://doi.org/10.1021/jm9602928>
151. Liu H, Li Z, Hall D, Liang P, Ma T. Sophia: a scalable stochastic second-order optimizer for language model pre-training. 2023; p [arXiv:2305.14342](https://arxiv.org/abs/2305.14342).
152. Opler LA, Feinberg SS (1991) The role of pimozide in clinical psychiatry: a review. *J Clin Psychiatry* 52(5):221–233
153. Fulop G, Phillips R, Shapiro A, Gomes J, Shapiro E, Nordlie J (1987) ECG changes during haloperidol and pimozide treatment of Tourette's disorder. *Am J Psychiatry* 144(5):673–675
154. Krähenbühl S, Sauter B, Kupferschmidt H, Krause M, Wyss PA, Meier PJ (1995) Reversible QT prolongation with torsades de pointes in a patient with pimozide intoxication. *Am J Med Sci* 309(6):315–316
155. Food; Administration, D.; Health, U. D. o.; Services, H. ORAP[®] (Pimozide) Tablets. 2008. https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/017473s041lbl.pdf.
156. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC50 data—a statistical analysis. *PLoS ONE* 8(4):e61007. <https://doi.org/10.1371/journal.pone.0061007>
157. Qar J, Galizzi J-P, Fosset M, Lazdunski M (1987) Receptors for diphenylbutylpiperidine neuroleptics in brain, cardiac, and smooth muscle membranes. Relationship with receptors for 1,4-dihydropyridines and phenylalkylamines and with Ca²⁺ channel blockade. *Eur J Pharmacol* 141(2):261–268. [https://doi.org/10.1016/0014-2999\(87\)90271-8](https://doi.org/10.1016/0014-2999(87)90271-8)
158. Tang B, Ewalt J, Ng H-L. Generative AI models for drug discovery. In: *Biophysical and computational tools in drug discovery*, Saxena AK, Ed. Springer International Publishing, 2021; pp. 221–243.
159. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. PyTorch: an imperative style, high-performance deep learning library. 2019; p [arXiv:1912.01703](https://arxiv.org/abs/1912.01703).
160. Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. 2019; p [arXiv:1903.02428](https://arxiv.org/abs/1903.02428).
161. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. 2019; p [arXiv:1907.10902](https://arxiv.org/abs/1907.10902).
162. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.