Supporting Information

CardioGenAI: A Machine Learning-Based Framework for Re-Engineering Drugs for Reduced hERG Liability

Gregory W. Kyro[1,*], Matthew T. Martin[2], Eric D. Watt[2], Victor S. Batista[1,*]

[1]Department of Chemistry, Yale University, New Haven, Connecticut 06511

[2]Drug Safety Research & Development, Pfizer Research & Development, Groton, Connecticut 06340

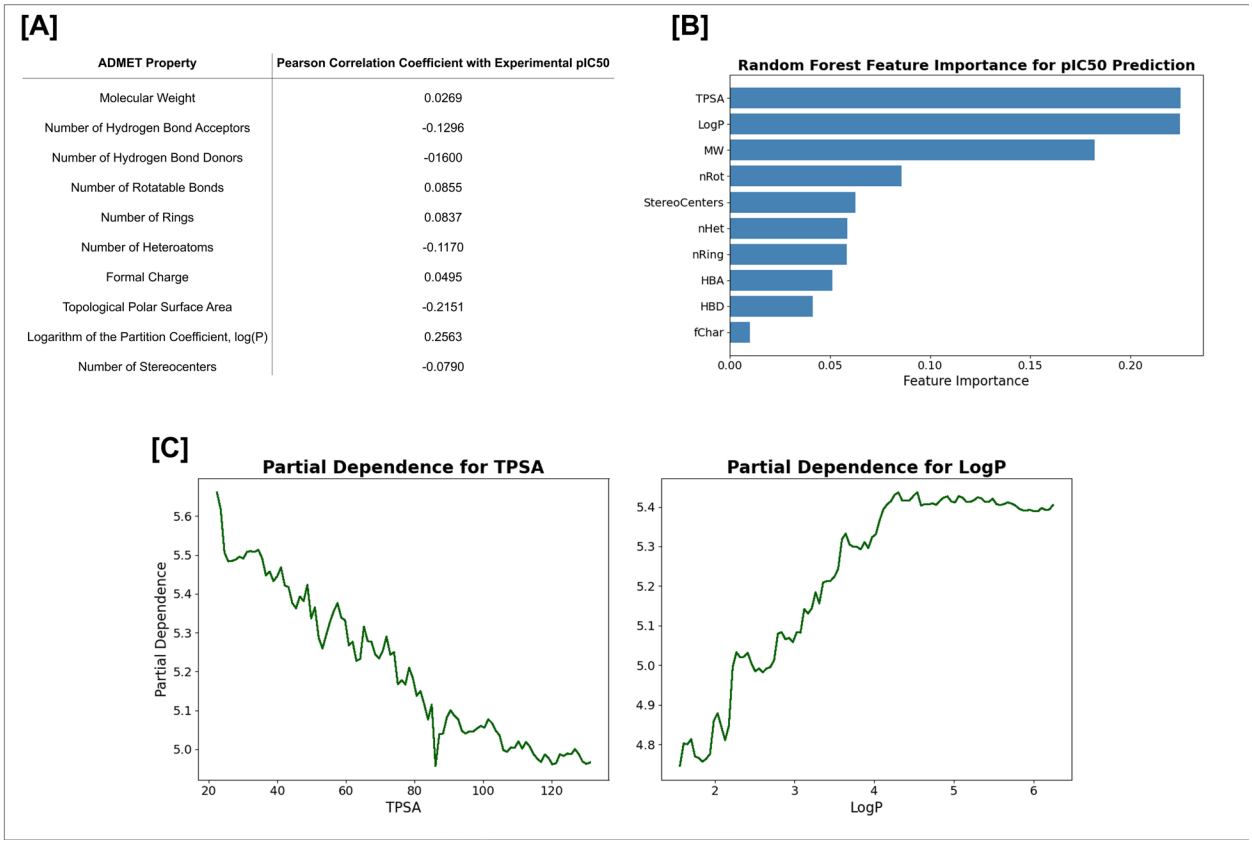

Figure S1. Analysis of the relationships between ADMET properties and experimental hERG channel pIC50 values. In [A], Pearson correlation is shown between each ADMET property and experimental pIC50 values obtained from the training set used. There are noteworthy correlations with pIC50 for topological polar surface area (TPSA) and LogP. A random forest model with 100 estimators was fit to the data to predict pIC50 values, and the importance of each feature was then deduced. In [B], the feature importance of each ADMET property is shown. In [C], the partial dependences for TPSA (Å^2) and LogP are shown. For LogP, there is an initial positive trend were an increase in LogP corresponds to an increase in PIC50 values up to a LogP value of approximately 4. For TPSA, as TPSA increases, pIC50 values generally decrease.

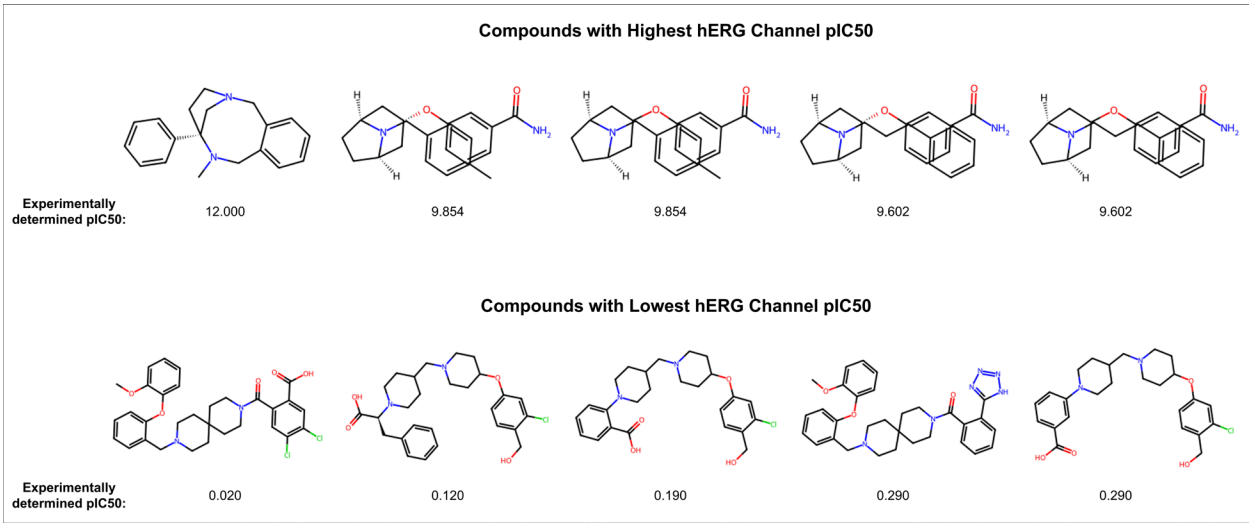

Figure S2. Display of the five compounds with highest hERG channel activity, as well as the five compounds with the lowest hERG channel activity of compounds in the hERG channel training set used. Each molecule is labeled with the corresponding experimentally determined hERG channel pIC50.

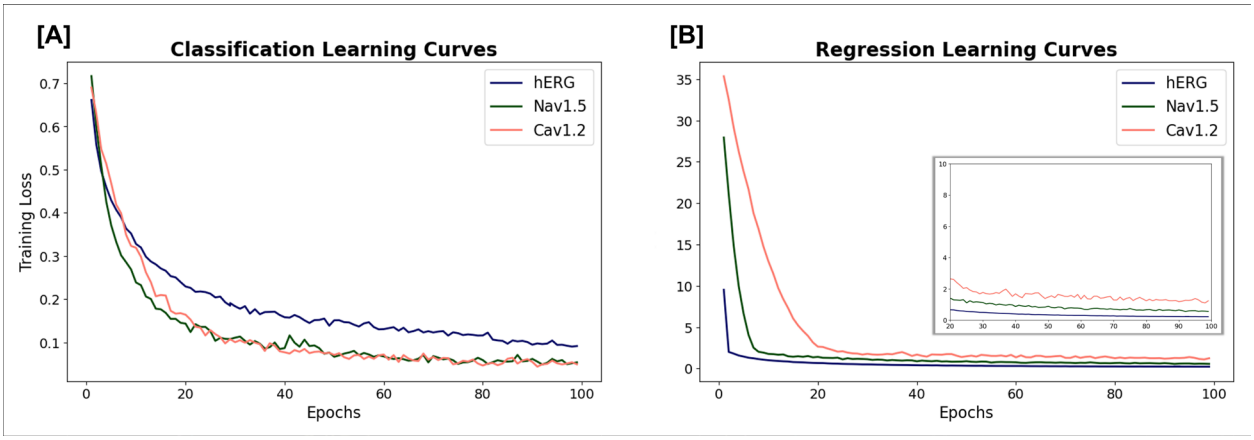

Figure S3. Learning curves for hERG, Nav1.5 and Cav1.2 cardiac ion channel [A] classification and [B] regression models. The classification models are trained with binary cross entropy loss and the AdamW optimizer for 100 epochs. Regression models are trained analogously but using mean squared error loss.

Table S4. Performance regarding each possible feature-representation combination for binary classification of hERG channel blockers.

Feature Representations	AC	SN	SP	F1	CCR	MCC
Transformer feature vector + Graph + Fingerprint	**83.5**	**86.2**	**80.3**	**85.1**	**83.2**	**66.7**
Transformer feature vector + Fingerprint	80.4	82.9	77.3	82.3	80.1	60.4
Transformer feature vector + Graph	80.0	82.9	76.4	81.9	79.6	59.5
Graph + Fingerprint	78.6	80.9	75.9	80.6	78.4	56.8
Transformer feature vector	77.7	83.3	70.9	80.4	77.1	54.9
Fingerprint	76.6	79.3	73.4	78.8	76.3	52.7
Graph	74.4	87.8	58.1	79.0	73.0	48.6

[a] The evaluation set used is that developed by Arab et al.; compounds in the evaluation 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.

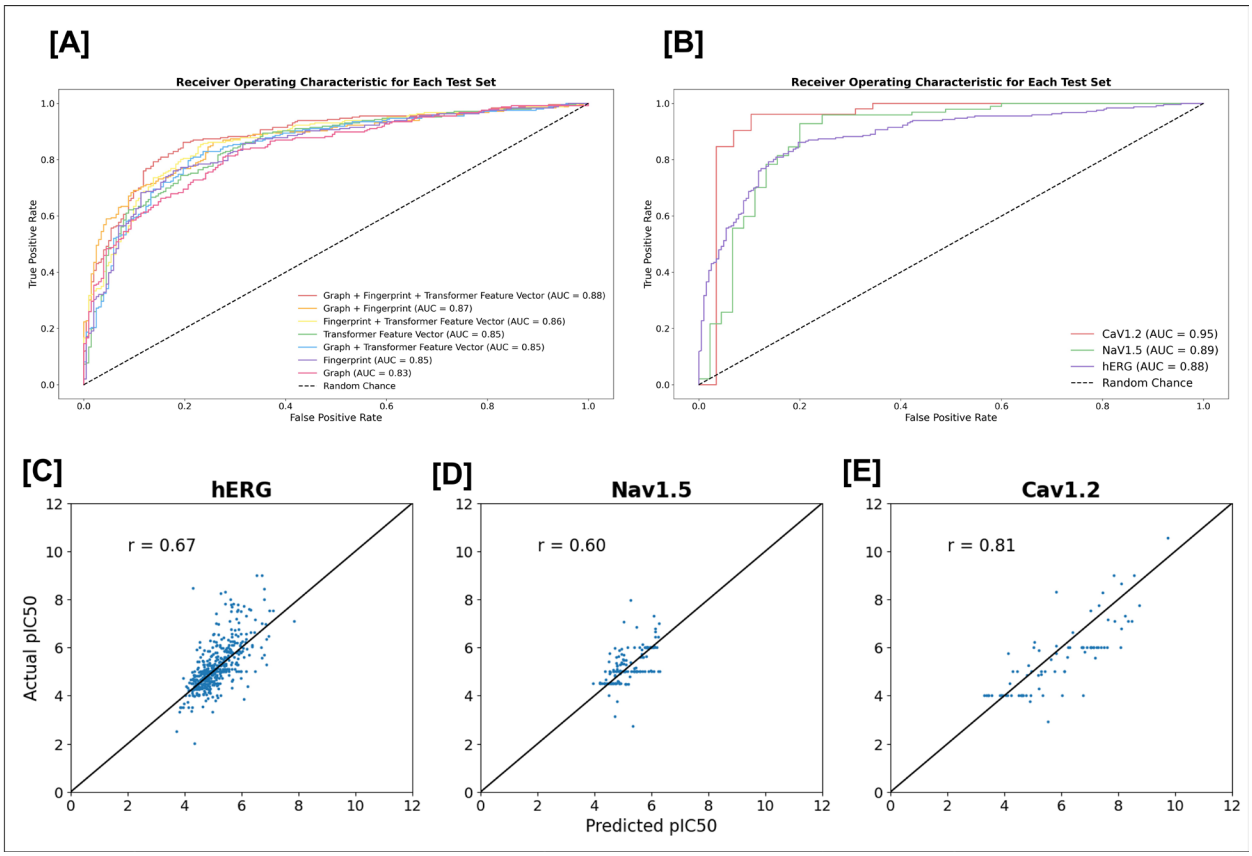

Figure S5. Evaluation of classification and regression models for cardiac ion channel blocker prediction. [A] Receiver operating characteristic (ROC) curve for evaluation on the hERG benchmark presented by Arab et al. Results regarding each feature-representation combination are shown with the corresponding area under the curve (AUC). [B] ROC curve for evaluation on the hERG, $Na_V1.5$ and $Ca_V1.2$ channel benchmarks. Scatter plots depicting actual pIC50 as a function of predicted pIC50, with the corresponding Pearson correlation coefficient (r), shown for evaluation on the [C] hERG, [D] $Na_V1.5$, and [E] $Ca_V1.2$ channel benchmark sets.

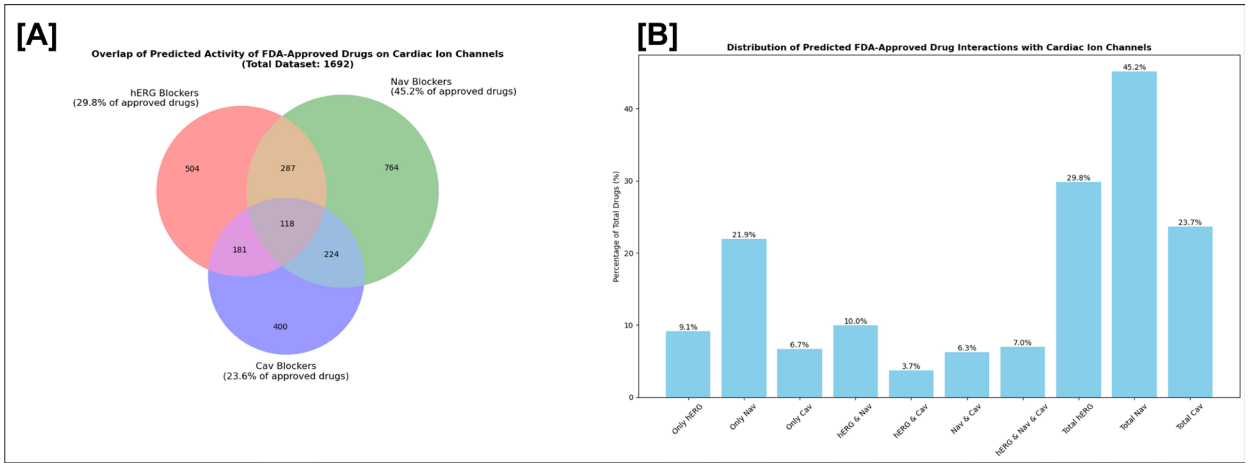

Figure S6. Application of cardiac ion channel classification models to a subset of FDA-approved drugs obtained from DrugCentral. A Venn diagram [A] and bar plot [B] are shown for the screening results of the 1692 total compounds.

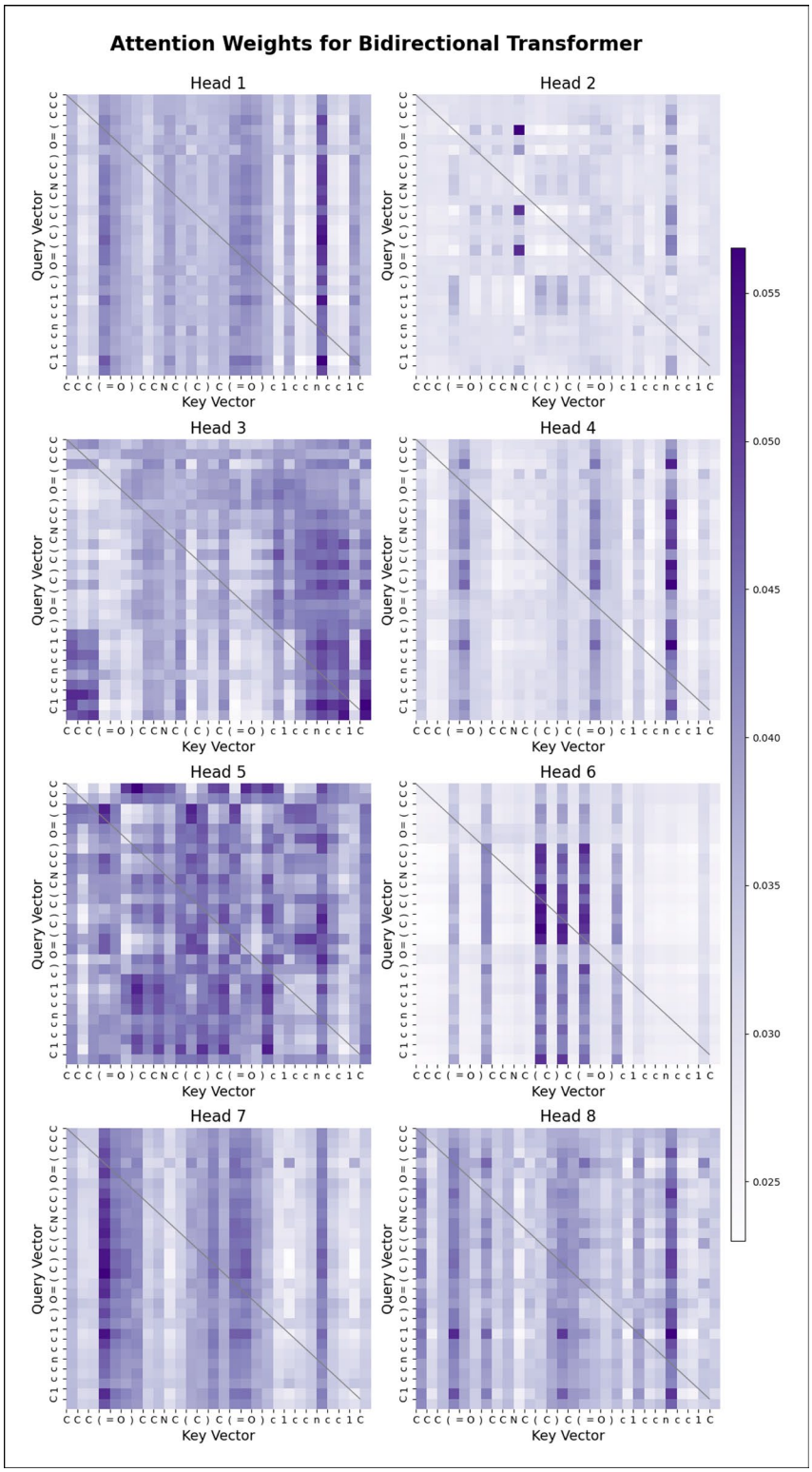

Figure S7. Representative attention maps depicting the distribution of attention weights extracted from the bidirectional transformer trained on SMILES strings, with each subplot corresponding to one of the eight heads in the model's attention layer. Weights are shown for the SMILES string: "CCC(=O)CCNC(C)C(=O)c1ccncc1C".

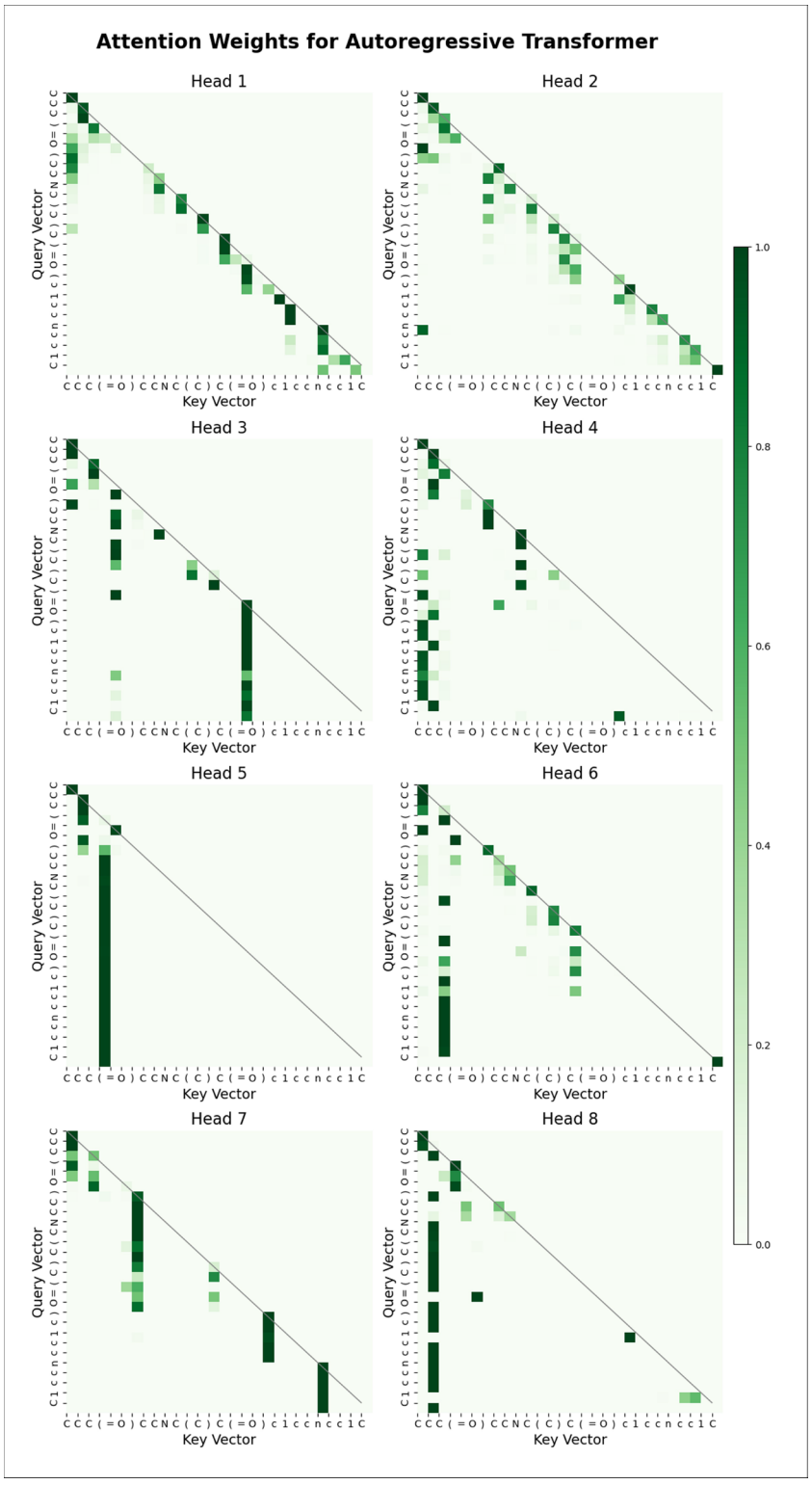

Figure S8. Representative attention maps depicting the distribution of attention weights extracted from the autoregressive transformer trained on SMILES strings, with each subplot corresponding to one of the eight heads in the model's attention layer. Weights are shown for the SMILES string: "CCC(=O)CCNC(C)C(=O)c1ccncc1C".

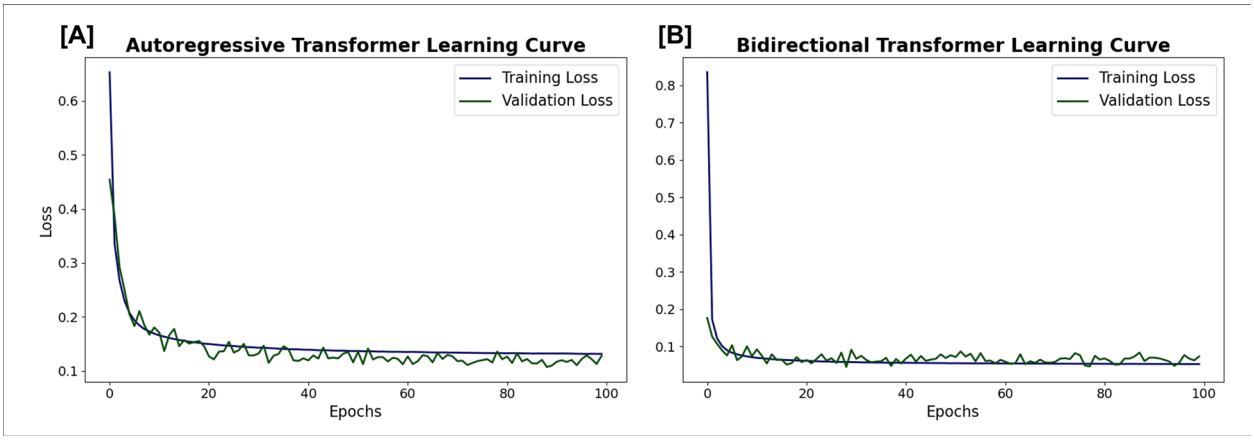

Figure S9. Learning curves for the [A] autoregressive transformer and [B] bidirectional transformer models. The autoregressive transformer is trained for next-token prediction, and the bidirectional transformer is trained for masked-token prediction. Both models are trained with cross entropy loss and the Sophia optimizer for 100 epochs.

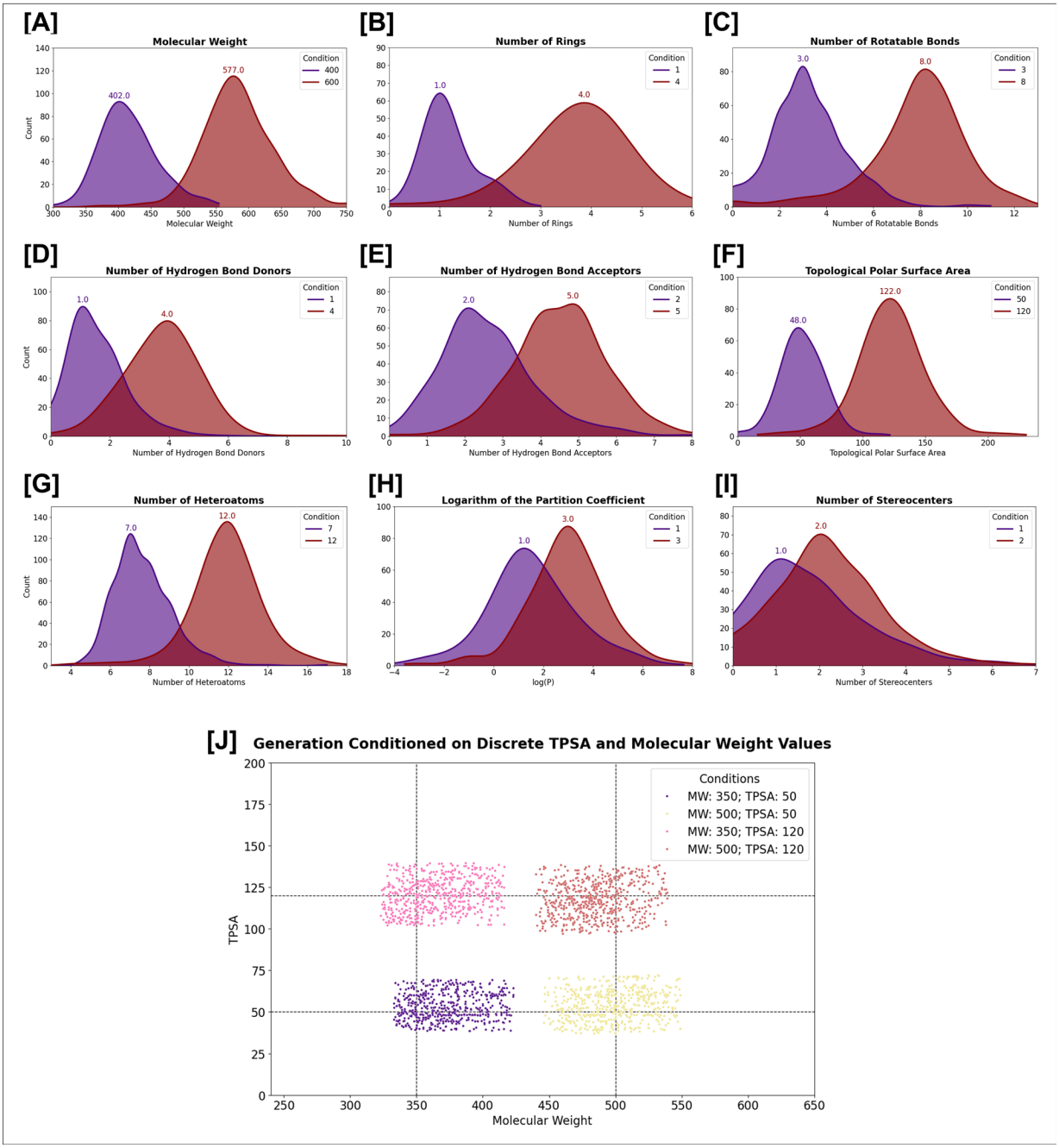

Figure S10. Probing the ability of the generative model to generate distributions of molecules with desired conditions. Results are shown for generations conditioned on different discrete values for [A] molecular weight ($\frac{g}{mol}$), [B] number of rings, [C] number of rotatable bonds, [D] number of hydrogen bond donors, [E] number of hydrogen bond acceptors, [F] topological polar surface area ($Å^2$), [G] number of heteroatoms, [H] LogP (logarithm of the partition coefficient), and [I] number of stereocenters. In [J], the generations are conditioned based on different value combinations of topological polar surface area (TPSA) and molecular weight. For each of the four pairs of conditions, values above/below two standard deviations greater/less than the mean value of each metric are excluded to emphasize the locations of the distribution means.

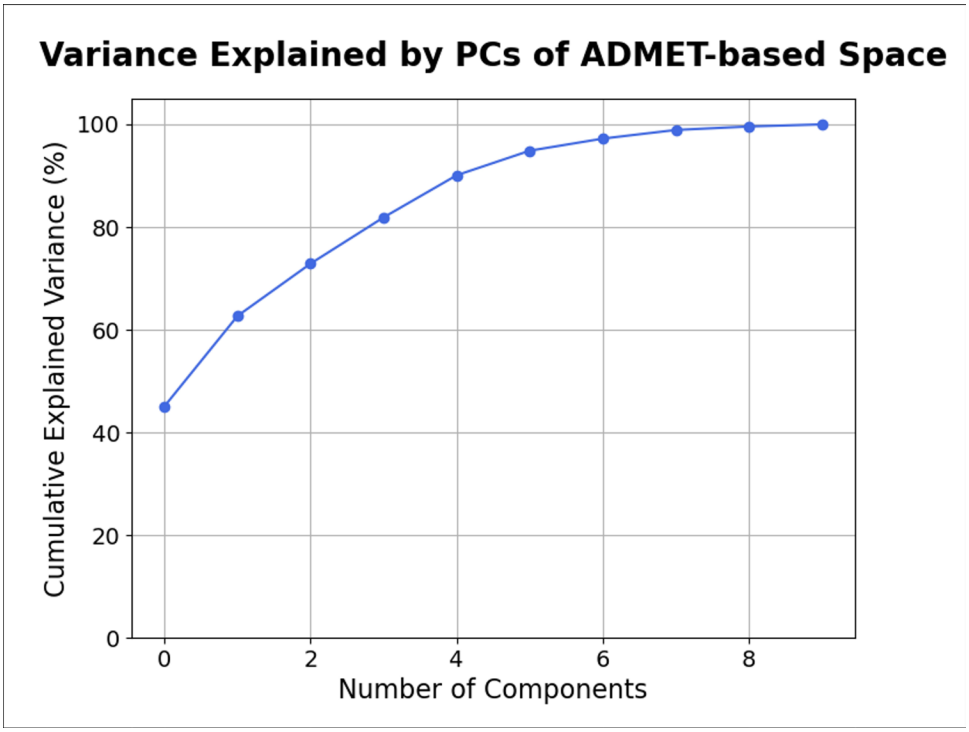

Figure S11. Cumulative variance as a function of principal component for principal component analysis (PCA) of ADMET-based chemical space. For the input cardiotoxic molecule (pimozide), the generated molecules (100 datapoints), and the molecules in the pretraining set for the autoregressive transformer-based generative model (approximately 5 million datapoints), ADMET properties are calculated, and PCA is performed to generate a lower-dimensional representation of chemical space.

Table S12. CardioGenAI methodology applied to pimozide, an FDA-approved antipsychotic drug that has a predicted hERG-channel pIC50 of 7.629, and is reported to cause hERG channel blockade-induced arrhythmias (Table 4). 100 molecules are generated, and among them is fluspirilene, a compound that belongs to the same class of drugs as pimozide and therefore has a similar primary therapeutic mode of action, but exhibits significantly less hERG-channel activity (5.785 pIC50). Included in the table are the five most similar generated compounds to pimozide in terms of cosine similarity between molecular descriptor vectors.

Similarity Rank	SMILES String	Cosine Similarity	Predicted pIC50
Input Molecule	O=c1[nH]c2ccccc2n1C1CCN(CCCC(c2ccc(F)cc2)c2ccc(F)cc2)CC1 (pimozide)	1.000	7.629
1	Fc1ccc(CCCN2CCc3c[nH]c(n3)C2)c(N2CCC(c3ccccc3)CC2)c1	0.980	5.367
2	O=C(Nc1ccccc1)N(CCCN1CCCC1)CC1(c2ccc(F)cc2)CCC1	0.977	5.904
3	O=C(NCc1ccc(F)cc1)N(CCCN1CCCC1)c1cccc2ccccc12	0.976	5.794
4	O=C1NCN(c2ccccc2)C12CCN(CCCC(c1ccc(F)cc1)c1ccc(F)cc1)CC2 (fluspirilene)	0.948	5.785
5	Fc1cccc(Cc2ncc(C3(N4CCC(Cc5ccccc5)CC4)CCOCC3)[nH]2)c1	0.946	5.201

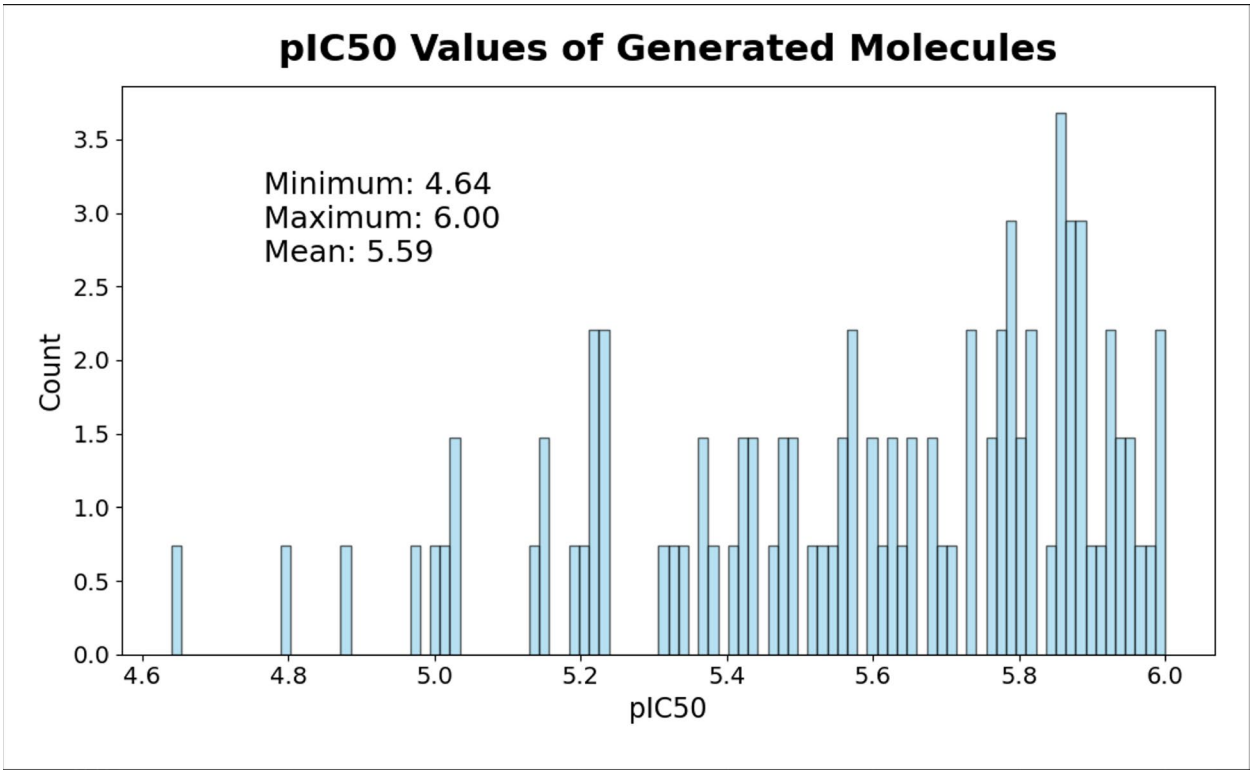

Figure S13. Predicted hERG channel pIC50 values for the 100 generated molecules conditioned on pimozide. The minimum (4.64), mean (5.59) and maximum (6.00) pIC50 values are shown.