Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

EXPLAINABLE NEURAL NETWORKS FOR CARDIOTOXICITY PREDICTION USING HERG, ION-CHANNEL, AND SUBSTRUCTURE FEATURES

Omar Khalid[1]*, Sara Nadeem[1], Bilal Farooq[2], Hina Saeed[1]

1. ***Department of Pharmaceutical AI Analytics, Faculty of Pharmacy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.***
2. ***Department of Intelligent Drug Systems, Faculty of Engineering, Qatar University, Doha, Qatar.***

ARTICLE INFO

Received:
02 November 2025
Received in revised form:
28 January 2026
Accepted:
01 February 2026
Available online:
28 February 2026

Keywords: Explainable AI, Cardiotoxicity prediction, hERG inhibition, Ion-channel profiling, Graph neural networks, Molecular substructures

ABSTRACT

Cardiac toxicity remains a major safety challenge in drug discovery, as pro-arrhythmic risk can emerge from complex interactions between chemical structure, hERG inhibition, and broader ion-channel pharmacology, with certain molecular substructures creating toxicophoric patterns that are not always captured by single-assay interpretation. While many computational cardiotoxicity models provide useful risk predictions, their limited interpretability reduces practical value for toxicologists and medicinal chemists, who need to understand why a compound is flagged and how it could be redesigned. To address this, an explainable neural network for cardiotoxicity prediction can integrate hERG information, multi-ion-channel profiles, and molecular substructure features, offering atom-level, substructure-level, and channel-level explanations for each prediction. Using a neural architecture such as graph attention or message passing to encode molecular graphs while incorporating ion-channel inhibition data, and applying post-hoc explanation methods like SHAP or integrated gradients, the model can decompose predictions into structural alerts and channel-specific contributions. Conceptually, it could identify a compound as pro-arrhythmic and clarify that the predicted risk is driven by hERG inhibition, sodium-channel activity, and highlighted substructures consistent with known cardiotoxic motifs. By linking predictive modeling to interpretable molecular and electrophysiological drivers, such an approach could enhance transparency, support safer drug design, and facilitate more informed decision-making in cardiac safety assessment.

To Cite This Article: Khalid O, Nadeem S, Farooq B, Saeed H. Explainable Neural Networks for Cardiotoxicity Prediction Using hERG, Ion-Channel, and Substructure Features. Pharmacophore. 2026;17(1):62-71. https://doi.org/10.51847/xpHg4tbqKZ

Introduction

Drug-induced cardiotoxicity remains a persistent cause of drug attrition because it can emerge from electrophysiological liabilities that are not evident from general pharmacological profiling alone [1]. The hERG potassium channel is central to cardiac safety assessment because inhibition of this channel can contribute to delayed repolarization and pro-arrhythmic risk [2]. However, focusing only on hERG may oversimplify risk, because compounds can also affect Nav1.5, Cav1.2, and other cardiac ion channels that modify the overall action-potential response [3]. A cardiotoxicity model therefore needs to represent hERG as a major but incomplete component of a broader mechanistic system [4].

Comprehensive in-vitro cardiac safety panels create an opportunity to move beyond single-target classification toward multi-channel interpretation [5]. Machine-learning models can synthesize heterogeneous channel data, chemical descriptors, and molecular graph representations into a unified risk estimate [6]. Such integration is especially important when a compound shows mixed activity across depolarizing and repolarizing currents, because these patterns may alter net pro-arrhythmic potential [7]. An explainable model should therefore report not only whether risk is predicted, but also which channels are expected to drive that prediction [8].

The regulatory and scientific movement toward more mechanistic cardiac safety assessment has encouraged computational models that connect ion-channel activity with downstream electrophysiological effects [7]. In-silico cardiac action-potential modeling can provide mechanistic context for interpreting multi-channel pharmacology, while data-driven models can learn

Corresponding Author: Omar Khalid; Department of Pharmaceutical AI Analytics, Faculty of Pharmacy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. E-mail: omar.khalid@gmail.com.

structure-risk associations from large chemical collections [9]. Explainability is essential in this setting because safety decisions require evidence that can be inspected, challenged, and linked to established toxicological reasoning [10]. A transparent neural network would be more useful than a black-box model when its explanations align with known cardiac pharmacology and chemical toxicophores [11].

This article describes an explainable neural network that integrates hERG inhibition, multi-ion-channel activity, and molecular substructure features for conceptual cardiotoxicity prediction [12]. The proposed model would combine a molecular graph encoder with channel-profile inputs, allowing it to represent both structural and mechanistic determinants of risk [13]. Explanation layers or post-hoc methods would then attribute each prediction to specific atoms, functional groups, and channel measurements rather than presenting only a risk label [14]. The goal is a model-oriented framework in which cardiotoxicity prediction becomes actionable for toxicologists and medicinal chemists [15].

Background

Drug-Induced Cardiotoxicity and Ion-Channel Pharmacology

Drug-induced cardiotoxicity often reflects disruption of cardiac electrophysiology, with hERG-mediated IKr inhibition serving as one of the most widely studied mechanisms [2]. Yet cardiac repolarization and conduction depend on multiple channels, so Nav1.5, Cav1.2, KvLQT1, and related currents can influence the final safety phenotype [3]. Multi-channel modeling is therefore consistent with the view that pro-arrhythmic risk should be interpreted as an integrated electrophysiological response rather than as a single molecular interaction [4]. A neural model that accepts both hERG and non-hERG channel inputs could better reflect this mechanistic complexity [5].

In-Silico Models for Cardiac Safety

In-silico cardiotoxicity models have evolved from structural alerts and descriptor-based QSAR toward deep learning approaches that encode chemical graphs and pharmacological profiles [1]. hERG prediction models have used diverse machine-learning strategies to classify blockers, estimate inhibition risk, and support early screening decisions [16]. Graph neural networks and attention-based systems extend this approach by learning molecular representations directly from atom-bond structures rather than relying only on fixed descriptors [17]. However, many high-capacity models still require explanation mechanisms before they can be used confidently in safety-critical decision-making [18].

Molecular Substructures and Toxicophores for Cardiotoxicity

Molecular substructures are central to cardiotoxicity interpretation because certain aromatic, lipophilic, and basic motifs can be associated with hERG binding and related pro-arrhythmic liabilities [19]. Fingerprint-based models can encode these motifs as explicit bits, while graph models can learn local atom environments that resemble known toxicophores [20]. Interpretable hERG models are especially valuable when they identify structural alerts that medicinal chemists can modify without losing desired pharmacological activity [10]. A substructure-aware neural network should therefore connect predicted risk to recognizable chemical groups rather than to abstract latent features alone [11].

Table 1 shows common molecular substructures associated with hERG-related cardiotoxicity and the modeling approaches that can capture them.

Table 1. Representative Molecular Substructures and Their Modeling Approaches in Cardiotoxicity Prediction

Substructure / Motif	Associated Cardiotoxic Risk	Modeling Approach	Interpretability Advantage
Aromatic rings	hERG binding, pro-arrhythmic liability	Fingerprint-based, Graph-based	Explicit bits or learned local environments highlight toxicophores
Lipophilic groups	hERG binding, membrane interaction	Fingerprint-based, Graph-based	Allows identification of hydrophobic alerts that can be modified
Basic amines	Ion channel interaction	Fingerprint-based, Graph-based	Structural alerts can guide chemical modifications without losing activity
Heterocyclic scaffolds	hERG and off-target liabilities	Graph-based neural networks	Captures complex atom environments not easily encoded in fingerprints

Explainable AI in Predictive Toxicology

Explainable AI in predictive toxicology aims to make model reasoning visible through feature attribution, attention weights, perturbation analysis, or counterfactual design suggestions [12]. SHAP values can attribute a prediction to molecular descriptors, channel measurements, or graph-derived features, while integrated gradients can trace the contribution of input features through a neural architecture [10]. Attention-based graph models can also highlight atoms or bonds that are influential for a prediction, although such attention maps should be interpreted carefully and validated against perturbation-based evidence [13]. The quality of an explanation depends on whether it is faithful to the model, chemically plausible, and useful for redesign decisions [14].

Current Limitations of Black-Box Cardiotoxicity Models

Black-box cardiotoxicity models may be accurate in a general predictive sense, but they provide limited support for safety scientists who need mechanistic justification [21]. A model that reports only a cardiotoxicity class cannot distinguish a primarily hERG-driven signal from a broader multi-channel liability, even though these cases may imply different follow-up experiments [5]. Similarly, a prediction without substructure attribution does not tell a chemist which chemical motif should be modified to reduce risk [15]. Current model development therefore needs stronger integration of multi-channel data, molecular explanations, and toxicological decision logic [8].

Model Development Overview

High-Level Prediction and Explanation Pipeline

The proposed pipeline begins with a molecular graph in which atoms and bonds are represented as structured inputs to a graph neural network [17]. A learned molecular embedding is then fused with supplemental channel features such as hERG, Nav1.5, and Cav1.2 activity, reflecting the multi-modal design used in recent cardiac safety modeling [5]. The network would output a conceptual cardiotoxicity risk score, after which attribution methods would identify the structural fragments and ion-channel features most responsible for the prediction [12]. This design allows the model to function as both a predictor and an explanation generator [11].

Core Input Features

Core input features would include atom type, bond order, aromaticity, formal charge, hydrogen-bonding patterns, and other graph-level chemical properties that help define local substructures [18]. hERG information would be included as a central channel feature, either as an inhibition label or as a continuous pharmacological measurement depending on data availability [2]. Additional ion-channel inputs such as Nav1.5 and Cav1.2 would provide complementary mechanistic context for distinguishing single-channel from multi-channel liability [3]. Optional physicochemical descriptors such as lipophilicity, polar surface area, and ionization-related properties could be added to support interpretable structure-risk reasoning [9].

Design Principles

The model should be multi-modal, combining molecular graph learning with explicitly encoded ion-channel measurements rather than treating cardiotoxicity as a purely structural classification task [6]. It should also support intrinsic or post-hoc explanation, using graph attention, SHAP-style feature attribution, or integrated-gradient analysis to localize risk drivers [13]. Because channel data may be incomplete, the architecture should handle missingness through indicator variables or masking rather than silently discarding compounds [3]. Finally, the output should translate model reasoning into safety-relevant language, distinguishing structural toxicophores from mechanistic channel contributions [4].

Table 2 maps each input and architectural component of the proposed model to its mechanistic cardiac-safety meaning, explanation output, and practical interpretive use.

Table 2. Mechanistic Input-to-Explanation Map for the Proposed Explainable Cardiotoxicity Neural Network

Input or model component	Primary representation in the framework	Cardiac-safety meaning	Explanation output enabled	Practical interpretation for toxicologists and medicinal chemists
Molecular graph	Atom nodes, bond edges, aromaticity, charge, heteroatom identity, local chemical neighborhoods	Captures structural determinants of cardiotoxicity that may not be fully represented by fixed descriptors	Atom-level and bond-level attribution maps	Identifies local chemical regions that the model associates with elevated or reduced predicted cardiac risk
hERG / IKr feature	Continuous inhibition value, blocker label, assay-derived activity class, or normalized pharmacological measurement	Represents a major repolarization-related liability associated with delayed cardiac repolarization	Channel-level hERG contribution score	Helps determine whether the prediction is primarily hERG-driven rather than broadly cardiotoxic
Nav1.5 / INa feature	Sodium-channel inhibition or activity feature, with missingness indicator when unavailable	Provides information about conduction-related electrophysiological liability	Channel-level sodium-current contribution	Helps distinguish repolarization-centered risk from broader multi-channel electrophysiological concern
Cav1.2 / ICa feature	Calcium-channel inhibition or activity feature, standardized across assay formats where possible	Adds mechanistic context for calcium-current effects that may modify net cardiac response	Channel-level calcium-current contribution	Supports interpretation of compounds with mixed ion-channel profiles rather than isolated hERG activity
Additional cardiac ion channels	Optional KvLQT1, late sodium current, or other available cardiac-channel features	Expands mechanistic coverage beyond the most common assay panel	Multi-channel attribution pattern	Clarifies whether predicted risk reflects a narrow or distributed electrophysiological signal

Substructure fingerprints	Explicit structural-alert bits, extended-connectivity fingerprints, toxicophore indicators	Encodes recognizable motifs associated with hERG binding or cardiotoxic liability	Substructure-level attribution	Converts abstract model reasoning into modifiable chemical features
Physicochemical descriptors	Lipophilicity, polar surface area, ionization state, hydrogen-bonding features	Provides interpretable chemical-property context for channel binding and permeability-related patterns	Descriptor-level contribution	Helps chemists understand whether predicted risk is linked to global molecular properties rather than a single fragment
Graph neural encoder	Message passing, graph attention, atom embeddings, molecular-level embedding	Learns latent structure-risk patterns from the molecular graph	Atom-importance and learned-neighborhood attribution	Reveals whether the model is focusing on chemically plausible regions of the molecule
Fusion layer	Gated or attention-based integration of molecular embedding and channel features	Combines structural and mechanistic pharmacology evidence	Separable structural versus channel contribution estimates	Shows whether prediction is driven by chemical structure, measured channel behavior, or both
Prediction head	Cardiotoxicity risk score or risk category	Produces the model's overall safety concern estimate	Prediction-level decomposition	Provides a risk signal that can be reviewed together with mechanistic and structural explanations
SHAP or feature attribution module	Contribution values for descriptors, channel features, and substructure bits	Quantifies which inputs increase or decrease the prediction	Ranked feature-contribution profile	Supports transparent review of why the model produced a specific prediction
Integrated-gradient or graph attribution module	Path-based or gradient-based contribution estimates for model inputs	Tests how input features influence the neural prediction	Atom-level, bond-level, or feature-level attribution	Adds an independent explanation view to reduce reliance on a single attribution method
Substructure masking or perturbation test	Removal, masking, or modification of highlighted fragments	Evaluates whether highlighted fragments are faithful drivers of the prediction	Explanation-faithfulness evidence	Helps avoid visually persuasive but model-unfaithful explanations
Narrative explanation layer	Structured natural-language summary of risk drivers	Translates model outputs into safety-review language	Human-readable explanation	Makes the result usable in project meetings, redesign discussions, and safety documentation
Human expert review	Toxicologist and medicinal-chemist inspection of prediction and explanation	Prevents automatic overreliance on computational outputs	Plausibility and actionability judgment	Frames the model as a transparent hypothesis generator rather than an experimental substitute

Data Sources and Feature Engineering

Compilation of Cardiotoxicity and Ion-Channel Datasets

A cardiotoxicity modeling resource would conceptually combine hERG measurements from public medicinal-chemistry and toxicology repositories with curated hERG-focused datasets used in prior prediction studies [21]. Multi-ion-channel information could be compiled from cardiac safety studies that include hERG, Nav1.5, Cav1.2, and related channels, allowing the model to learn mechanistic profiles rather than isolated endpoints [3]. Broader cardiotoxicity labels could be linked from clinical safety annotations, drug-label evidence, and pharmacovigilance-derived outcomes when those labels are defined consistently [9]. Dataset harmonization should preserve assay provenance because patch-clamp, fluorescence, and other assay formats may not represent identical biological signals [4].

Encoding Molecular Substructures and hERG Fingerprints

Molecular substructures can be encoded through extended-connectivity fingerprints, graph representations, and explicit toxicophore indicators so that the model has both learned and human-readable chemical inputs [20]. hERG-oriented fingerprints are useful because they can represent structural motifs that appear repeatedly among blockers and nonblockers [16]. A graph encoder can complement these fixed fingerprints by learning atom-level neighborhoods that may capture more flexible or context-dependent toxicophore patterns [17]. Including explicit substructure bits alongside learned embeddings also makes attribution easier to interpret because a feature can correspond to a recognizable functional group [10].

Standardization of Ion-Channel Data

Ion-channel activity data should be standardized before modeling because measurements may be reported in different concentration units, assay conditions, and categorical formats [3]. Harmonized hERG, Nav1.5, and Cav1.2 values could be transformed into comparable feature representations, while missing measurements could be encoded with separate indicators to avoid confusing absence of data with absence of activity [6]. Assay variability should be treated as an interpretive limitation, particularly when combining public datasets with different experimental protocols [21]. Feature engineering should therefore retain enough metadata for downstream explanations to distinguish strong mechanistic evidence from uncertain or incomplete channel information [5].

Explainable Neural Network Architecture

Graph-Based Encoder

A graph-based encoder would process the molecule as an atom-bond network, learning representations that reflect local chemical environments and longer-range structural context [18]. A graph attention network could assign different weights to neighboring atoms, while a message-passing neural network could iteratively update atom embeddings based on bonded interactions [16]. These learned embeddings would allow the model to represent aromatic systems, charged amines, heteroatoms, and other motifs relevant to hERG or broader cardiac safety [19]. Because attention values alone may not guarantee faithful explanations, the encoder should be paired with attribution methods that can test whether highlighted atoms truly influence the prediction [14].

Fusion with Ion-Channel Features

The fusion layer would combine the molecular embedding with normalized channel features so that structural and pharmacological evidence contribute jointly to the risk estimate [5]. A gated or attention-based fusion mechanism could allow the model to emphasize hERG for compounds whose risk appears primarily repolarization-related, while giving more weight to Nav1.5 or Cav1.2 when the channel profile suggests broader electrophysiological involvement [3]. Such a design would conceptually align with multi-ion-channel safety platforms that interpret cardiac risk as a combined mechanistic profile rather than a single endpoint [4]. The fused representation should remain decomposable so that later explanations can separate structural contributions from channel-specific contributions [8].

Output and Explainability Layer

The output head would produce a conceptual cardiotoxicity risk estimate without requiring the model to be presented as a definitive experimental substitute [1]. Post-hoc SHAP analysis could decompose this risk into contributions from channel features, descriptors, and substructure indicators, while graph-level attribution could highlight influential atoms and bonds [10]. Integrated-gradient-style explanations could provide a complementary view of how input features contribute through the neural network, especially when the architecture uses continuous molecular or channel embeddings [12]. The final explanation should therefore report what the model predicts, which mechanistic channel signals support the prediction, and which molecular fragments appear most responsible [11].

Figure 1 presents the proposed explainable neural-network architecture linking molecular graph encoding, hERG and multi-ion-channel fusion, substructure attribution, and human-reviewed cardiotoxicity decision support.

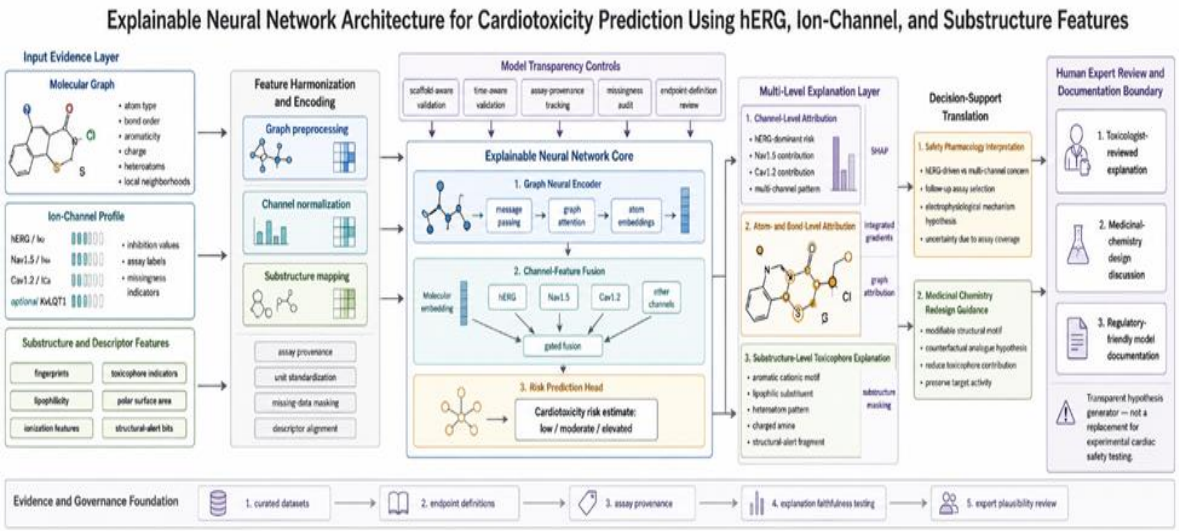

Figure 1. Explainable Neural Network Architecture for Cardiotoxicity Prediction Using hERG, Ion-Channel, and Substructure Features.

Linking Predictions to Mechanistic Safety Markers and Substructures

Decomposing the Prediction by Ion Channel

A decomposable cardiotoxicity prediction should show whether the model's reasoning is dominated by hERG inhibition or by a broader pattern of ion-channel activity [5]. For a compound with strong hERG evidence but limited non-hERG activity, the explanation would be expected to emphasize repolarization liability as the principal mechanistic driver [22]. In contrast, a compound with combined hERG, Nav1.5, and Cav1.2 activity could be described as having a multi-channel safety concern, consistent with models that integrate complementary channel profiles [3, 6]. This channel-level decomposition would help safety scientists decide whether follow-up work should focus on hERG selectivity, broader cardiac electrophysiology, or both [4].

Substructure Attribution and Toxicophore Identification

Substructure attribution should convert atom-level importance into chemically meaningful functional-group explanations, because medicinal chemists rarely act on isolated atom scores alone [10]. A graph-based model could highlight aromatic cationic motifs, lipophilic substituents, or heteroatom patterns that resemble known hERG-associated toxicophores [19]. Attention-based hERG models suggest that localized structural reasoning can be aligned with molecular features relevant to blockade, provided that attention is validated against complementary attribution methods [22]. The preferred explanation would therefore connect the predicted channel signal to a modifiable structural motif rather than merely stating that the compound is risky [11].

Global Pattern of Cardiotoxicity Drivers

Across a compound library, global explanation summaries could reveal recurring structural and mechanistic drivers of predicted cardiotoxicity [14]. Such summaries might show that particular scaffold families are repeatedly associated with hERG-dominant predictions, while others show broader multi-channel patterns involving sodium or calcium channels [3]. Graph attention and message-passing models can support this kind of library-level interpretation because they learn recurring molecular environments rather than only fixed descriptor combinations [17, 18]. Global explanations should be used to guide series-wide safety optimization while preserving compound-level explanations for individual design decisions [15].

Counterfactual Explanations for Safety Redesign

Counterfactual explanations can make a cardiotoxicity model more actionable by asking how a prediction would change if a suspected toxicophore were modified [14]. Instead of reporting a numerical redesign claim, the model could state conceptually that replacing a highly basic terminal amine with a less cationic group would be expected to reduce hERG-driven risk if the rest of the molecule remains pharmacologically acceptable [10]. This reasoning aligns with explainable multi-ion-channel assessment platforms, where the goal is to connect a proposed structural change with a plausible reduction in mechanistic liability [4]. Counterfactual outputs should always be treated as design hypotheses that require experimental confirmation rather than as definitive safety conclusions [8].

Explainability Methods for Cardiac Safety Assessment

Interactive Visualization for Toxicologists

An interactive visualization layer would overlay molecular attributions on a two-dimensional structure while separately displaying channel-level contributions [11]. This interface could allow toxicologists to inspect whether the highlighted atoms correspond to recognizable hERG-associated motifs and whether the channel profile supports the predicted mechanism [12]. For multi-channel cases, a visual comparison of hERG, Nav1.5, Cav1.2, and related features would clarify whether the model is reasoning from one dominant signal or from a broader electrophysiological pattern [5]. Such visualization would be most useful when paired with provenance information about assay type and data completeness [21].

Narrative Explanation Generation

A narrative explanation should translate feature attributions into safety-relevant language that can be reviewed by toxicologists and medicinal chemists [15]. Rather than stating a numerical probability, the explanation could say that the compound is predicted to have elevated cardiotoxicity concern because the model assigns strong importance to hERG activity and to a cationic aromatic substructure consistent with known blocker patterns [23]. If non-hERG channel features also contribute, the narrative should identify them as complementary mechanistic evidence rather than treating the prediction as a simple hERG classification [3]. This kind of structured natural-language output would make the model easier to use in project discussions and safety review documents [8].

Benchmarking Explanation Faithfulness

Explanation faithfulness should be evaluated by testing whether the highlighted atoms, substructures, or channel features are actually influential for the model's prediction [14]. A perturbation-based assessment could conceptually remove or mask the most important features and examine whether the model's risk interpretation changes in the expected direction, without relying on reported numerical performance claims [12]. Substructure-masking approaches are especially relevant because they can test whether chemically intuitive fragments genuinely drive molecular property predictions [14]. Faithfulness assessment should also compare highlighted motifs with known hERG and cardiotoxicity structural knowledge to avoid explanations that are visually appealing but mechanistically weak [10].

Regulatory-Friendly Model Documentation

Regulatory-friendly documentation should describe the model architecture, data provenance, endpoint definitions, feature engineering, and explanation method in a form that can be audited [7]. For cardiac safety, this documentation should clarify how hERG and non-hERG channels are represented and how model explanations relate to mechanistic safety markers [4]. It should also distinguish model-generated hypotheses from experimental evidence, because an explainable prediction does not replace patch-clamp, action-potential, or other confirmatory studies [5]. Clear documentation would support transparent communication among computational modelers, safety pharmacologists, project teams, and regulatory reviewers [8].

Integration Into Medicinal Chemistry And Safety Pharmacology

Early-Stage De-Risking of Lead Series

During lead optimization, an explainable cardiotoxicity model could help prioritize compounds by combining predicted cardiac safety concern with interpretable structural guidance [24]. If a lead series repeatedly shows attributions around the same cationic or lipophilic motif, chemists could explore analogues that reduce that motif's contribution while preserving target activity [10]. hERG-specific tools and broader cardiotoxicity prediction models both suggest that early computational triage can support safer design when predictions are interpreted with chemical context [25, 26]. The model should therefore function as a decision-support system that proposes risk hypotheses rather than as an automatic compound rejection tool [15].

Supporting Investigative Toxicology and Regulatory Submissions

In investigative toxicology, the model's explanation could help frame why a compound should undergo additional cardiac safety experiments or why a structural modification may be justified [8]. A report might describe that the predicted concern arises from both a hERG-associated substructure and a multi-channel inhibition profile, linking the computational signal to plausible electrophysiological mechanisms [4, 5]. For regulatory submissions, explainable outputs would be most useful when they are accompanied by documented data sources, model assumptions, and limitations in channel coverage [7]. This framing would allow computational evidence to complement, rather than replace, established safety pharmacology studies [24].

Evaluation Strategy

Predictive Performance

Predictive evaluation should compare the explainable neural network with descriptor-based QSAR models, hERG-focused classifiers, and black-box graph neural networks using scaffold-aware and time-aware validation concepts [1]. The purpose would not be to claim a specific performance level, but to determine whether the model maintains useful discrimination while adding interpretable reasoning [21]. hERG-focused benchmarks and cardiotoxicity models provide a basis for conceptual comparison across model families, especially when endpoint definitions and data provenance are harmonized [16, 25]. The evaluation should also consider whether multi-channel fusion improves mechanistic plausibility compared with models that rely only on molecular structure or hERG labels [3].

Table 3 provides a deployment-readiness framework for evaluating whether the proposed explainable cardiotoxicity model is predictive, mechanistically plausible, faithful, stable, and actionable.

Table 3. Evaluation and Deployment Readiness Framework for Explainable Cardiotoxicity Prediction

Evaluation domain	Core question	Recommended assessment approach	Evidence needed before practical use	Main failure mode addressed	Decision-support implication
Predictive discrimination	Does the model distinguish compounds with higher versus lower cardiotoxicity concern?	Compare against descriptor-based QSAR, hERG-only classifiers, and black-box graph models using harmonized endpoints	Scaffold-aware and time-aware validation results; clear endpoint definitions; comparison across model families	A model appears innovative but does not improve or maintain useful predictive performance	Determines whether the explainable architecture is worth using as a screening-support tool
Scaffold generalization	Does the model remain useful for chemically novel series?	Use scaffold split validation and external chemical-series testing	Performance across unseen scaffolds, not only random train-test splits	Overfitting to familiar chemical families	Supports early discovery use where new analogues may differ from training compounds
Time-aware robustness	Does the model generalize to compounds developed or tested after the training period?	Train on earlier data and evaluate on later compounds or later assay releases	Temporal validation design with documented data cutoff	Inflated performance from retrospective data leakage	Increases confidence that the model can support future project decisions
hERG-specific interpretability	Can the model separate hERG-dominant risk from other risk patterns?	Decompose predictions into hERG and non-hERG channel contributions	Channel-level attribution reports aligned with known hERG pharmacology	Treating all cardiotoxicity predictions as simple hERG classification	Helps select whether follow-up should prioritize hERG testing or broader electrophysiology

Multi-channel mechanistic plausibility	Do Nav1.5, Cav1.2, and other channels contribute in biologically plausible ways?	Review attribution patterns across compounds with known multi-channel activity	Expert pharmacology review of channel-contribution profiles	Spurious channel attributions caused by incomplete or biased assay data	Supports mechanistic interpretation rather than purely statistical classification
Substructure explanation quality	Are highlighted molecular fragments chemically meaningful and redesign-relevant?	Compare atom and substructure attributions with known toxicophores and medicinal-chemistry judgment	Expert review of highlighted motifs; consistency across related analogues	Highlighting visually plausible but chemically irrelevant atoms	Determines whether the explanation can guide analogue design
Explanation faithfulness	Are the highlighted features genuinely influential for the model's prediction?	Use substructure masking, perturbation tests, integrated gradients, and attribution comparison	Prediction changes after masking or modifying highly attributed features	Explanations that look convincing but do not reflect model reasoning	Prevents overinterpretation of attractive but unfaithful explanation graphics
Explanation stability	Are explanations consistent under small input or modeling changes?	Evaluate attribution consistency across random seeds, nearby analogues, and alternative explanation methods	Stability metrics or structured qualitative comparison	Fragile explanations that change without meaningful chemical differences	Supports reliable use in compound-series discussions
Assay-provenance transparency	Does the model distinguish strong channel evidence from uncertain or heterogeneous assay evidence?	Track assay type, source, concentration format, and missingness indicators	Provenance metadata linked to each channel feature	Treating heterogeneous assay measurements as equivalent	Helps toxicologists judge whether a prediction rests on strong or weak experimental evidence
Missing-data handling	Does the model avoid confusing missing channel data with absence of activity?	Use explicit missingness indicators, masking, and sensitivity analysis	Documentation of missing-data logic and performance under incomplete profiles	False reassurance when non-hERG channel measurements are unavailable	Improves interpretation of compounds with partial ion-channel panels
Counterfactual usefulness	Do suggested structural changes form plausible medicinal-chemistry hypotheses?	Generate and review counterfactual modification scenarios without claiming experimental certainty	Expert assessment of synthetic feasibility, target-activity preservation, and toxicophore reduction	Unrealistic or unsafe redesign suggestions	Positions counterfactuals as hypotheses for chemist review, not automatic optimization instructions
Human decision impact	Do explanations improve project decisions compared with risk scores alone?	Retrospective case review and prospective user studies with toxicologists and chemists	Evidence that explanations support assay selection, analogue prioritization, or clearer safety discussion	Accurate predictions that remain unused because they are not actionable	Establishes whether the model improves real decision quality
Documentation readiness	Is the model sufficiently transparent for audit, review, and regulated communication?	Prepare model cards, data-provenance summaries, endpoint definitions, and explanation-method documentation	Auditable documentation of architecture, inputs, limitations, and intended use	Unclear model assumptions or unsupported regulatory claims	Enables responsible communication with project teams and regulatory reviewers
Deployment boundary	Is the model used as a decision-support tool rather than an experimental replacement?	Define intended use, non-use cases, uncertainty statements, and required confirmatory testing	Explicit governance language and human-review requirement	Automation bias or replacement of cardiac safety assays	Preserves the model's role as a transparent hypothesis generator

Explanation Quality and Actionability

Explanation quality should be assessed through chemical plausibility, toxicological relevance, and actionability for compound redesign [10]. Expert reviewers could judge whether highlighted atoms and substructures correspond to known hERG or cardiotoxicity motifs and whether the proposed channel-level explanation is consistent with cardiac safety pharmacology [4, 19]. Graph explanation methods such as substructure masking, attention analysis, and attribution comparison can provide complementary checks on whether the explanation is stable and faithful to the model [14, 17]. Actionability should be defined by whether the explanation suggests a realistic medicinal-chemistry hypothesis rather than only a generic warning [15].

Impact on Compound Design

The impact of the model on compound design should be evaluated conceptually through retrospective and prospective decision-support studies [24]. Retrospective analysis could compare model explanations with later medicinal-chemistry changes in historical lead series, asking whether the highlighted substructures match motifs that chemists eventually modified [15]. Prospective use could examine whether project teams find the explanations useful for prioritizing analogues, selecting follow-up assays, and discussing cardiac safety risk earlier in discovery [26]. Such evaluation should emphasize decision quality and interpretive value rather than presenting unsupported claims about reduced attrition or guaranteed safety improvement.

Limitations

Incomplete Ion-Channel Coverage and Dynamic Effects

A key limitation is that even a multi-channel neural network may include only a subset of the cardiac currents that shape pro-arrhythmic risk [3]. Time-dependent block, metabolite effects, trafficking disruption, chronic cardiomyocyte stress, and non-electrophysiological toxicity may not be captured by hERG, Nav1.5, and Cav1.2 inputs alone [7]. Public datasets also differ in assay technology and curation quality, which may influence both prediction and explanation [21]. Therefore, model outputs should be interpreted as risk hypotheses that require mechanistic follow-up rather than as complete representations of cardiac biology [5].

Dependence on Explanation Method

Another limitation is that post-hoc explanations can vary depending on the attribution method, feature representation, and background distribution used for comparison [12]. Attention maps, SHAP values, integrated gradients, and perturbation-based explanations may emphasize different aspects of the same prediction, so agreement among methods should be examined before drawing strong mechanistic conclusions [13, 14]. Users may also overinterpret highlighted fragments as causal toxicophores, even when the model has learned correlations from biased data [10]. Training and documentation are therefore necessary so that explanations support careful expert reasoning rather than false confidence [8].

Conclusion

An explainable neural network for cardiotoxicity prediction can provide a structured way to connect molecular structure, ion-channel pharmacology, and cardiac safety reasoning. By integrating hERG information with broader channel profiles and substructure features, the model can move beyond simple blocker classification toward mechanistically informed risk interpretation. Its main value is not only the predicted risk category but also the explanation of why that risk is expected. This makes the framework suitable for decision support in early drug discovery and safety assessment.

The strength of the approach lies in its ability to produce explanations at multiple levels of biological and chemical meaning. Channel-level attributions can indicate whether a prediction is hERG-dominant or reflects broader electrophysiological concern. Atom-level and substructure-level explanations can identify motifs that may be responsible for the predicted signal. Together, these outputs can give medicinal chemists and toxicologists a clearer basis for redesign, assay selection, and project discussion.

Important challenges remain before such a system could be used confidently in a regulated or project-critical setting. Ion-channel coverage may be incomplete, assay variability may affect interpretation, and post-hoc explanations may be sensitive to modeling choices. Prospective validation in real drug-discovery programs would be needed to determine whether the explanations genuinely improve decisions. The model should therefore be viewed as a transparent hypothesis generator rather than as a replacement for experimental cardiac safety testing.

Progress in explainable cardiac safety modeling will require collaboration among computational scientists, medicinal chemists, toxicologists, electrophysiologists, and regulators. Computational teams can build transparent architectures and robust attribution workflows, while domain experts can test whether explanations are biologically plausible and practically useful. Regulators can help define documentation standards and evaluation expectations for model-supported safety decisions. A shared evaluation framework would make explainable AI more credible, reproducible, and actionable for cardiotoxicity prediction.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, et al. Deep learning-based prediction of drug-induced cardiotoxicity. J Chem Inf Model. 2019;59(3):1073-84.
2. Ryu JY, Lee MY, Lee JH, Lee BH, Oh KS. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics. 2020;36(10):3049-55.
3. Arab I, Egghe K, Laukens K, Chen K, Barakat K, Bittremieux W. Benchmarking of small molecule feature representations for hERG, Nav1.5, and Cav1.2 cardiotoxicity prediction. J Chem Inf Model. 2023;64(7):2515-27.
4. Gambacorta N, Mastrolorito F, Togo MV, Amenduni V, Mele M, Liantonio A, et al. CUPID: a free drug discovery platform for the explainable multi-ion channel assessment of cardiotoxicity. Eur J Med Chem. 2025;290:117575.
5. Chen Z, Li N, Zhang P, Li Y, Li X. CardioDPi: An explainable deep-learning model for identifying cardiotoxic chemicals targeting hERG, Cav1.2, and Nav1.5 channels. J Hazard Mater. 2024;474:134724.
6. Arab I, Laukens K, Bittremieux W. Semisupervised learning to boost hERG, Nav1.5, and Cav1.2 cardiac ion channel toxicity prediction by mining a large unlabeled small molecule data set. J Chem Inf Model. 2024;64(16):6410-20.
7. Fuadah YN, Qauli AI, Marcellinus A, Pramudito MA, Lim KM. Machine learning approach to evaluate TdP risk of drugs using cardiac electrophysiological model including inter-individual variability. Front Physiol. 2023;14:1266084.
8. Çelik FK, Doğan S, Karaduman G. Drug-induced torsadogenicity prediction model: An explainable machine learning-driven quantitative structure-toxicity relationship approach. Comput Biol Med. 2024;182:109209.
9. Delre P, Lavado GJ, Lamanna G, Saviano M, Roncaglioni A, Benfenati E, et al. Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. Front Pharmacol. 2022;13:951083.
10. Sanches IH, Braga RC, Alves VM, Andrade CH. Enhancing hERG risk assessment with interpretable classificatory and regression models. Chem Res Toxicol. 2024;37(6):910-22.
11. Yang T, Ding X, McMichael E, Pun FW, Aliper A, Ren F, et al. AttenhERG: a reliable and interpretable graph neural network framework for predicting hERG channel blockers. J Cheminform. 2024;16(1):143.
12. Kim H, Park M, Lee I, Nam H. BayeshERG: a robust, reliable and interpretable deep learning model for predicting hERG channel blockers. Brief Bioinform. 2022;23(4):bbac211.
13. Vinh T, Nguyen L, Trinh QH, Nguyen-Vo TH, Nguyen BP. Predicting cardiotoxicity of molecules using attention-based graph neural networks. J Chem Inf Model. 2024;64(6):1816-27.
14. Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nat Commun. 2023;14(1):2585.
15. Agarwal D, Sharma A, Garg P. Graph-Based Classification with GNN-Explainer for Predicting Cardiac Toxicity Associated with Multi-Ion Channel Blockers. Chem Res Toxicol. 2026.
16. Shan M, Jiang C, Chen J, Qin LP, Qin JJ, Cheng G. Predicting hERG channel blockers with directed message passing neural networks. RSC Adv. 2022;12(6):3423-30.
17. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2019;63(16):8749-60.
18. Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform. 2021;13(1):12.
19. Creanza TM, Delre P, Ancona N, Lentini G, Saviano M, Mangiatordi GF. Structure-based prediction of hERG-related cardiotoxicity: a benchmark study. J Chem Inf Model. 2021;61(9):4758-70.
20. Zhang X, Mao J, Wei M, Qi Y, Zhang JZ. HergSPred: accurate classification of hERG blockers/nonblockers with machine-learning models. J Chem Inf Model. 2022;62(8):1830-9.
21. Siramshetty VB, Nguyen DT, Martinez NJ, Southall NT, Simeonov A, Zakharov AV. Critical assessment of artificial intelligence methods for prediction of hERG channel inhibition in the "big data" era. J Chem Inf Model. 2020;60(12):6007-19.
22. Kim H, Nam H. hERG-Att: Self-attention-based deep neural network for predicting hERG blockers. Comput Biol Chem. 2020;87:107286.
23. Karim A, Lee M, Balle T, Sattar A. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. J Cheminform. 2021;13(1):60.
24. Liu K, Cui H, Yu X, Li W, Han W. Predicting cardiotoxicity in drug development: a deep learning approach. J Pharm Anal. 2025:101263.
25. Falcón-Cano G, Morales-Helguera A, Lambert H, Cabrera-Pérez MÁ, Molina C. hERG toxicity prediction in early drug discovery using extreme gradient boosting and isometric stratified ensemble mapping. Sci Rep. 2025;15(1):15585.
26. Tran-Nguyen VK, Randriharimanamizara UF, Taboureau O. HERGAI: an artificial intelligence tool for structure-based prediction of hERG inhibitors. J Cheminform. 2025;17(1):110.