scientific reports

OPEN

hERG toxicity prediction in early drug discovery using extreme gradient boosting and isometric stratified ensemble mapping

Gabriela Falcón-Cano[1], Aliuska Morales-Helguera[1], Heather Lambert[1], Miguel-Ángel Cabrera-Pérez[2] & Christophe Molina[1]✉

Blockade of the human Ether-à-go-go Related Gene (hERG) potassium channel by small molecules can prolong the QT interval, leading to fatal cardiotoxicity. Numerous drugs have been withdrawn from the market due to cardiac side effects, underscoring the need for early identification of hERG toxicity. Despite several classification machine learning (ML) models having been developed to this end, robustness, class imbalance, and interpretability are still challenges. Using the largest public database of hERG inhibition, this work integrates eXtreme Gradient Boosting (XGBoost) with Isometric Stratified Ensemble (ISE) mapping (XGB + ISE map) to enhance hERG prediction. An XGBoost consensus model was developed using balanced training sets and diverse variable subsets, resulting in robust models less affected by class imbalance. The model demonstrated competitive predictive performance, achieving a balance between sensitivity (SE = 0.83) and specificity (SP = 0.90) through exhaustive validation. ISE mapping estimated the model applicability domain and improved prediction confidence evaluation and compound selection by stratifying data. Refined variable selection procedures enhanced model interpretability. Variable importance analysis highlights key molecular determinants (peoe_VSA8, ESOL, SdssC, MaxssO, nRNR2, MATS1i, nRNHR, nRNH2) associated with hERG inhibition. The XGB + ISE map strategy provides an effective approach to identifying promising molecules in drug discovery campaigns with reduced hERG inhibition risk.

Keywords Machine learning, Ensemble methods, Imbalanced dataset, Applicability domain, Variable selection

Ether-à-go-go (EAG) proteins are potassium channels found in muscle tissues, the brain, endocrine cells, and the heart. The EAG-related gene (ERG) subfamily includes three distinct isoforms: Kv11.1, Kv11.2, and Kv11.3. The human isoform Kv11.1, or human Ether-à-go-go Related Gene (hERG), is crucial in drug development due to its role in cardiac toxicity. Although different mechanisms have been associated with QT-related arrhythmias like Torsades de Pointes (TdP), such as mitochondrial toxicity, inhibition of voltage-gated calcium channels, and hERG trafficking inhibition, blockade of the hERG potassium channel is considered one of the major reasons for drug-induced TdP[1–3]. Several marketed drugs, have been withdrawn due to toxicity associated with hERG blockade[4]. The high sensitivity of the hERG channel to inhibition by a wide variety of structurally diverse molecules reinforces the importance of early evaluation of hERG blockade during drug discovery[5,6].

Traditionally, anti-target studies have relied on a combination of in vitro and in vivo methods. However, ethical concerns and the high cost associated with in vivo methods have demanded the development of viable alternatives[7], aligning with the 4R principles: reduce, refine, replace and responsibility[8,9]. In vitro methods, like radioligand binding and patch clamp techniques, assess hERG liability, but are costly and time-consuming. In vivo hERG assays, while also expensive and time-consuming, further suffer from low throughput, making them almost impractical for evaluating large datasets of compounds in the early stages of drug discovery[10]. In this sense, the Food and Drug Administration has recently proposed a major initiative via the Comprehensive In Vitro Pro-Arrhythmia Assay program that integrates in silico models with in vitro data from engineered human cells and stem cell-derived cardiomyocytes to estimate cardiotoxicity in early drug development[11].

[1]PIKAÏROS, S.A, 31650 Saint Orens de Gameville, France. [2]Departamento de Ciencias Farmacéuticas, Facultad de Ciencias, Universidad Católica del Norte, Angamos, 0610 Antofagasta, Chile. ✉email: christophe.molina@pikairos.com

Computational toxicity prediction has become a valuable tool for assessing cardiotoxicity and eliminating molecules likely to fail in preclinical or clinical studies. Recent reviews, such as the one by Tran et al.[7], highlight that most hERG prediction models rely on public, diverse data from sources like ChEMBL, PubChem, and BindingDB, often combining data from various cell lines (e.g., CHO, Hek293).

The largest public database for predicting hERG inhibition is derived from the study by Sato et al.[2]. In their research, the authors meticulously compiled and integrated hERG-related data from ChEMBL, PubChem, GOSTAR, and hERG Central, resulting in a structurally diverse database. For the pivotal endpoint, many researchers have chosen to use half maximal inhibitory concentration (IC_{50}) to binarize hERG activity. However, the classification of compounds as inhibitors or non-inhibitors varies widely due to differing IC_{50} thresholds, which can range from 1 to 40 µM[7]. In classification models, the selection of these thresholds is closely tied to the issue of data imbalance, a recognized challenge in machine learning (ML). To address this, researchers in the field of hERG prediction have employed various methods, such as simple data manipulation techniques (e.g., Synthetic Minority Over-sampling TEchnique (SMOTE), random oversampling and random undersampling). More sophisticated approaches include weighted models and probability threshold optimization based on classification metrics sensitive to class imbalance[12–14].

In terms of molecular descriptors, most studies rely on classical physicochemical descriptors[12,15–17] and/or molecular fingerprints[6,13,18–22]. Additionally, some studies explore graph structures to uncover correlations with hERG inhibition[23,24]. Modeling algorithms predominantly include deep learning (DL) neural networks (NN), such as Recurrent NN, Graph Convolutional NN, and Generative Adversarial NN[6,19–22,24], as well as ML methods like Random Forest[12,16,25], Support Vector Machine (SVM)[13] and eXtreme Gradient Boosting (XGBoost)[17,26]. Model comparison is often challenging due to the use of proprietary datasets and the lack of publicly available code for reproducing results or evaluating new external sets. Regarding model performance, statistical results from different approaches tend to show minimal variation in global performance indicators, such as accuracy (ACC, ranging from 0.80 to 0.95) and Area Under the Curve (AUC, ranging from 0.80 to 0.95) that is calculated from Receiver Operating Characteristic (ROC) curves. Although the published models demonstrate similar ACC, they should be analyzed using statistical metrics that reflect class imbalance, as these metrics reveal substantial differences in their ability to predict inhibitor and non-inhibitor samples. Nevertheless, a standardized benchmarking test set remains essential for making meaningful model comparisons. Additionally, many current models lack interpretability and explainability, which poses a challenge in modeling hERG, particularly given the critical implications for drug safety[7,16,27].

This work implements a refined ML algorithm integrating recursive methods for variable selection, XGBoost and Isometric Stratified Ensemble (ISE) mapping to enhance hERG prediction. XGBoost, a gradient boosting machine variant, has demonstrated superior predictive performance in hERG channel inhibition modeling, outperforming other ML algorithms[17]. Its ability to handle imbalanced datasets and maintain robustness across diverse chemical spaces[28] makes it particularly suitable for cardiotoxicity screening in early drug discovery. By integrating XGBoost with recursive feature selection and ISE mapping, this study aims to optimize predictive performance, expand the applicability domain (AD), and enhance model interpretability. Inspired by our previous work on antileishmanial activity and colloidal aggregation[29,30], this strategy leverages the ISE Map methodology, adapting it for hERG inhibition prediction by incorporating model-agnostic interpretability techniques and enhancing screening power. This leads to an advanced strategy to integrate hERG inhibition risk assessment into drug discovery campaigns to select promising compounds with reduced hERG liability.

Methods

General workflow

The Konstanz Information Miner (KNIME) open-source software version 4.7.8 (available free of charge at https://www.knime.com/download) and the KNIME Python Integration were used for the development and automatization of the Quantitative Structure–Activity Relationship (QSAR) methodology. Installation and usage instructions are available on the KNIME website. Figure 1 summarizes the general modeling pipeline described in the following sections.

Dataset

To develop the classification model, the largest public dataset with experimental values of hERG inhibitory activity was used[2,13]. This dataset made up of 291,219 molecules, comes from different in vitro assays and consists of 9,890 hERG inhibitors (molecules with $IC_{50} \leq 10$ µM; for molecules without IC_{50} information, any molecule with % of inhibition ≥ 50% at 10 µM). The remaining 281,329 molecules are considered non-inhibitors. Although the authors had previously outlined a detailed procedure for dataset curation, an additional curation protocol was implemented as described in our previous paper[31]. In brief, the initial stage involved the elimination of structures with erroneous representations, such as unusual valences, and structures containing d-block elements. Major fragments were retained for hydrates and inorganic salts during the manipulation of unconnected structures. The second stage focused on normalizing charges and bonds, achieved through the standardization of specific chemotypes, tautomeric forms, and the neutralization of charges. To perform these normalizations, the rules published by Kamel et al. were used[32]. These rules were manually drawn, saved as RxN files, and applied in the RDKit Chemical Transformation node within the KNIME platform. The third and final stage encompassed the removal of duplicates and the curation of experimental data. In this step, 3D features were excluded before generating International Chemical Identifier (InChI) codes/InChIKeys to identify duplicate molecules. Any molecule with an inconsistent class label was excluded from the dataset.

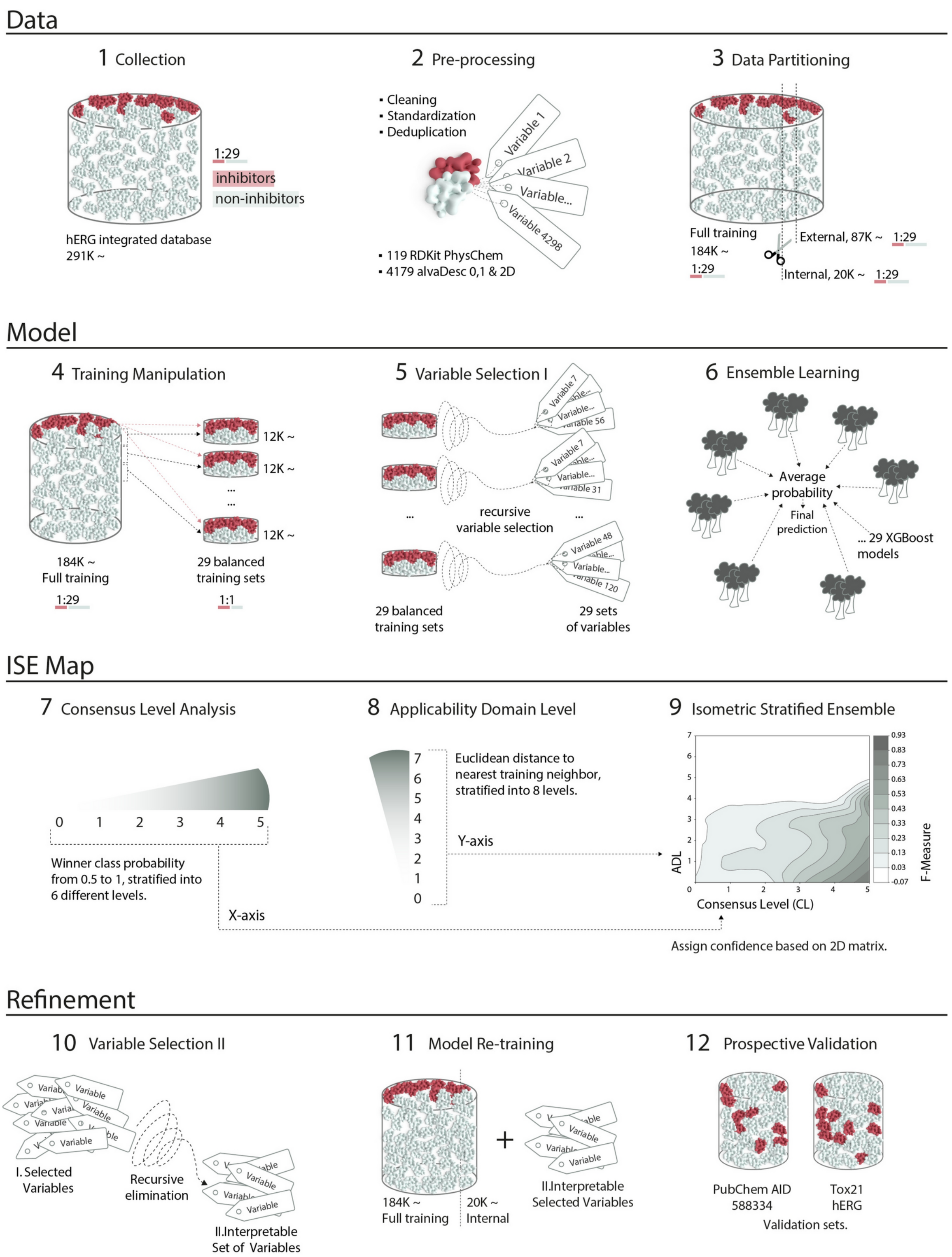


Fig. 1. General overview of the modeling pipeline.

Molecular descriptors

The molecular structure parameterization scheme was based solely on the 2D structure. Therefore, stereoisomers of the same molecule were considered duplicates and removed in the previous steps. Various 2D molecular representations were investigated to assess their significance in predicting hERG inhibition. Physicochemical properties, Molecular Operating Environment (MOE) and Kappa type descriptors, as well as Morgan, Feat Morgan, and MACCS fingerprints, were computed using the KNIME RDKit plugin[33]. Other groups of physicochemical descriptors were calculated using alvaDesc provided by Alvascience SRL [34,35]. These belong to the following families: Constitutional Indices, Ring Descriptors, Topological Indices, Walk and Path Counts, Connectivity Indices, Information Indices, 2D Matrix-Based Descriptors, 2D Autocorrelations, Burden Eigenvalues, P_VSA-like Descriptors, Extended Topochemical Atom (ETA) indices, Functional Group Counts, and Molecular Properties. The variable selection procedure was performed during the model building process as explained in the Variable Selection I subsection.

Data partitioning

The test set employed in the study by Ogura et al. was subtracted from the original dataset to act as an external test set (ET I), enabling a fair comparison with other models[13]. ET I represents 30% of the hERG dataset. From the remaining 70%, referred to here as the Modeling set, an internal test set was created by randomly selecting 10% of the samples. The remaining 90% formed the full training set. Given the size of our modeling set (> 200,000 compounds), allocating 90% of the data to training allows the model to capture complex patterns more effectively, while the 10% test set remains large enough to provide a reliable performance estimate. Our previous studies support this decision[29,30], which used this ratio for large datasets. The full training set, internal test set, and ET I set are all highly imbalanced, with the same imbalance ratio (see Table 1 of the Results and Discussion section). This guarantees an unbiased evaluation of model performance during training and testing phases.

Prospective validation

To ensure comprehensive validation of the model and conduct a thorough analysis of its performance, two additional external datasets were assembled. This validation was performed after constructing and optimizing the model. The entire modeling set was utilized to make predictions on each new external dataset, ensuring the removal of any overlaps between the new external data and the modeling set. Below, we provide a detailed description of the two External Validation Sets:

- *External Validation Set I (EV I).* The PubChem AID 588,834 consists of a 1536-well cell-based assay for high-throughput screening (HTS), specifically quantitative HTS (qHTS), designed to evaluate in vitro hERG channel blockage as a measure of cardiotoxicity with small molecules[36]. This assay employs the U2OS cell line derived from human osteosarcoma, using the FluxORTM thallium flux approach, and the same National Center for Advancing Translational Science (NCATS) laboratory. This dataset consists of a total of 5,381 data points taken from version 2.1 of PubChem AID 588,834. Among these data points, there are 664 compounds identified as HERG inhibitors, with an IC_{50} value of less than or equal to 10 μM, indicating their potential inhibitory activity. It is worth noting that the Ogura dataset [2] used in this study contains 2,688 chemicals taken from the same source, but from version 1.1. For this reason, EV-I is partially included in the original modeling set, and its overlap was removed from the modeling set before the validation process. This dataset is a valuable resource for model validation since it corresponds to an HTS assay.
- *External Validation Set II (EV II).* The second dataset is part of the US federal Tox21 program. The NCATS applied a qHTS approach to screen the Tox21 library of 9667 chemicals at 15 concentrations in triplicate to quantify IC_{50} values for hERG activity in the U2OS cell line using a thallium flux assay platform[37]. Missing SMILES (Simplified Molecular Input Line Entry System) and inconclusive samples were removed. IC_{50} values were used to define inhibitors ($IC_{50} \leq 10$ μM).

hERG Dataset	Class	Input Molecules	Output Molecules	Data Loss (%)	Imbalance Ratio	Murcko Scaffolds
Modeling	Total	203,853	203,425	0.21	28.42	71,831
	hERG	6,923	6,915			
	non-hERG	196,930	196,510			
ET I	Total	87,366	87,306	0.07	28.46	42,630
	hERG	2,967	2,964			
	non-hERG	84,399	84,342			
Prospective Validation						
EV I	Total	5,381	3,457	35.76	10.68	1,355
	hERG	664	296			
	non-hERG	4,107	3,157			
EV II	Total	9,667	6,509	33.1	24.53	1,883
	hERG	371	255			
	non-hERG	8,858	6,254			

Table 1. Summary of the data curation pipeline for training, test and validation datasets. The modeling dataset includes both training and internal test data. ET I is the External Test Set (Ogura's Test Set)[13], while EV I and EV II are external validation sets extracted from PubChem assay AID 588,334 [36] and Tox21 hERG data[37], respectively.

Balanced training sets

To prevent training bias toward non-inhibitory samples (the majority class) in the highly imbalanced full training set, multiple balanced training sets were generated. Specifically, the majority class was randomly sampled and divided into *R* subsets, where *R* represents the imbalance ratio of 29 (the number of non-inhibitor compounds divided by the number of inhibitor compounds). This process resulted in 29 non-inhibitor subsets, each of which was combined with the full set of inhibitor compounds to generate 29 balanced training sets. A more detailed discussion of this strategy can be found in our previous paper[29].

Quality analysis of the different datasets

A critical analysis of the different datasets was conducted following data curation to assess both experimental and chemical quality. Regarding experimental quality, the number of structures with inconsistent classification and missing data labels were determined. Subsequently, the overlap between the different datasets was analyzed. To assess the inter-dataset consistency, an overlap analysis was performed to quantify the degree of agreement in their hERG activity labels. In addition, an intra-dataset analysis was also conducted, focusing on molecules with two or more experimental bioactivity measurements (inhibitor/non-inhibitor) to identify agreement and potential inconsistencies within individual datasets. To measure the structural diversity of the compounds, the number of Murcko frameworks was counted, as reported by Langdon et al.[38].

Variable selection I

A molecular descriptor selection procedure, named Recursive Decorrelated Variable Selection (RVS), was applied to each of the 29 balanced training sets. The procedure determined the most important variables using a DT algorithm, followed by a recursive selection to remove redundant variables (see Variable selection II subsection below). Specifically, before variable selection, each descriptor column was duplicated, and their descriptor values were shuffled. This way, artificial random columns were appended to the original descriptor columns. To evaluate variable importance, the DT model was used to compute the number of occurrences of each original and artificial descriptor. A variable was selected if the ratio between the number of occurrences of the original variable and the number of occurrences of the counterpart artificial variable exceeded the threshold of 1.5. This means that the original variable was at least 1.5 times more important than the artificial counterpart. Afterwards, the Pearson correlation coefficient was calculated between the selected original variables. Through a recursive loop, the number of variables was reduced until no more correlated pairs existed above the defined Pearson correlation threshold ($r = 0.6$)[29–31].

Ensemble model

The strategy employed in this work relied on the construction of an ad hoc ensemble consisting of homogeneous base models trained on balanced training sets and validated on highly imbalanced data[29,30].

The ensemble model was composed of 29 XGBoost[39] base models, which were trained using 29 different balanced training sets (see Fig. 1 and Balanced Training Sets subsection). The parameters of the algorithm were configured as follows: booster = 'gbtree', learning_rate = 0.3, max_depth = 10, subsample = 1.0, colsample_bytree = 1.0, scale_pos_weight = 1.0, objective = 'binary', num_boost_round = 101. Finally, the XGBoost ensemble was first tested on the internal test set and then evaluated with ET I.

Applicability domain level

The concept of Applicability Domain Level (ADL) was introduced in our recent paper to refer to stratified intervals of molecular distance[30]. ADL is not a direct quantification of the Applicability Domain (AD) itself, but rather a practical framework based on "Molecule Distance Levels" that enables the evaluation of model performance relative to molecular novelty. This operational interpretation facilitates a more nuanced reliability analysis across a gradient of similarity to the training compounds.

To establish ADLs, Euclidean distances were calculated from each test set molecule to its nearest neighbor in the full training set. Prior to this calculation, molecular descriptors were normalized. Based on these nearest neighbor distances, eight ADLs were created, numbered from 0 to 7.

Consensus level

A consensus level (CL) was used to set the confidence level or probability of class labeling. Each one of the 29 XGBoost base models was used to predict the new data. Once the predictions of base models were completed, the probabilities for the inhibitor class were averaged, and from this, the confidence score or maximum probability of belonging to the winning class was calculated. Based on this probability, samples were categorized into subsets called 'strata', ranked from the highest to the lowest possible probability of belonging to the winning class. Winner class probability values ranging from 0.5 to 1 were linearly distributed into six different CLs numbered from zero to five, where zero represents the stratum with the lowest confidence and five represents the stratum with the highest confidence.

ISE map for estimating AD and prediction confidence

The ISE method integrates the CL and ADL into a 2D discrete map, the ISE map. In this map, compounds are positioned according to their CL (horizontal axis) and ADL (vertical axis) values. This grid organization enhances compound selection and ranking by considering both probability and AD criteria. With known statistics and compound cardinality in each cell, new compounds can be projected onto the ISE map in order to predict their statistics, and make informed decisions about compound selection and triage[30].

The ISE map can be analyzed from either a partial or an incremental perspective. In a partial ISE map, each sample is assigned to only one stratum, allowing for the calculation of individual statistics per stratum. In contrast, in an incremental ISE map, a sample can belong to multiple cumulative categories. Both approaches, partial and incremental, enhance the screening power of the method.

Variable selection II

Starting with the initial subset of key descriptors identified through the previous RVS, a second variable selection method named Recursive Variable Elimination (RVE) was applied to determine the minimum number of descriptors needed to preserve model performance. The process began with an initial model encompassing the previously selected descriptors, which were sorted in descending order based on their importance derived from the RVS method. Following this, an iterative process of descriptor elimination was implemented. In each iteration, the descriptor with the lowest importance was identified and removed. After the removal of each variable, its impact on the internal test set prediction was evaluated to assess whether its removal negatively affected or maintained the model performance. At each step, the importance of the remaining variables was recalculated, and the process was repeated, continuously removing the next least important variable.

With the iterative results of this method, a statistical comparison across iterations was conducted to determine the minimum number of variables needed for a simpler and more interpretable model, while still maintaining its performance. The Friedman non-parametric test was applied first, and when significant differences were found, a Dunn-Bonferroni test was performed for further verification.

Performance measures and evaluation

Several statistical measures were computed, encompassing Recall or Sensitivity (SE), Specificity (SP), Positive Predictive Value or Precision (PPV), Negative Predictive Value (NPV), F-measure, ACC, Balanced Accuracy (BACC) and Cohen's Kappa (CK). Additionally, the G-Mean and Matthews correlation coefficient (MCC) were determined. The AUC was derived from the ROC curve. Mathematical definitions for all statistics can be found in Table S1 in Supplementary Information (SI) 1.

To address the problem of model evaluation based on imbalanced data, F-measure, G-Mean, and MCC statistics were employed. The entire modeling pipeline was independently executed three times with different random seeds while consistently maintaining the same internal and ET I sets. The mean and standard deviation of the performance measures for the models were calculated and recorded for all training, internal test, and external test sets. Throughout the study, the ET I remained untouched during model training. The internal test set served multiple purposes, including the primary evaluation of model performance, the definition of isometric strata levels based on CL and ADL, and the selection of the final list of descriptors through the RVE procedure.

Statistical analysis and model retraining

The modeling and validation process were repeated independently three times using different random seeds. This approach facilitated the provision of mean and standard deviation values for all statistical parameters and frequency metrics related to the importance of the selected descriptors. Only those descriptors that appeared in all three runs and in at least 9 of the 29 base models (1/3 of the base models), were retained. At the end of this procedure, the first list of the most important variables was obtained (Selected Variables I). Subsequently, the RVE method was applied to the Selected Variables I set, resulting in a second list, the Interpretable Set of Variables (refer to Fig. 1). This led to having two additional models, one trained with Selected Variables I and the second trained using the Interpretable Set of Variables.

Results and discussion

Data analysis

The reliability and generalizability of results in QSAR modeling heavily depend on the quality and comprehensiveness of the dataset used. Table 1 offers a comprehensive overview of the datasets used for model development and prospective validation after the completion of data curation.

The Modeling set, used as the primary data source for training and internal testing, covers a diverse spectrum of structural classes and chemical functionalities. This diversity ensured that the Modeling set accurately represented the broad chemical space relevant to the biological endpoint under study[2,13]. The final ratio of non-hERG/hERG inhibitor compounds (IR = 29:1) highlights a considerable class imbalance, with 96.6% of the curated dataset biased toward non-inhibitor molecules. In silico modeling with such an imbalance typically exhibits weak predictive performance for the minority class (hERG inhibitors). To address this challenge, a strategy was implemented based on multiple balanced data sampling (MBDS), as described in prior published studies[29,30]. To implement MBDS, 29 base models were created, each using a different balanced training set with all inhibitors and an equal number of randomly sampled non-inhibitors. Internal and external sets (ET I, EV I and EV II) exhibited a bias toward non-inhibitor molecules, with imbalance ratios ranging from 11 to 25 (Table 1). The curated datasets, along with their set label (training, internal test, and ET I) are available in SI 2, while the curated EV I and EV II datasets are available in SI 3.

An overlap analysis between hERG datasets is provided in Fig. 2. Note that the dataset sourced from PubChem AID 588,334 (EV I) shows the highest number of duplicate molecules and intra-set classification agreement. The near-perfect agreements (> 90% in all cases) obtained for all datasets confirmed that, despite originating from different sources and experimental conditions, these datasets yielded closely aligned results for overlapping molecules, indicating generally favorable agreement in biological activity.

Figure S1 in SI 1 illustrates cumulative frequency plots of Murcko scaffolds across each dataset. Acyclic molecules were excluded from the analysis of Murcko scaffolds, which apply exclusively to ring-containing structures[38,41]. Nonetheless, the percentage of acyclic molecules for each dataset was relatively consistent and low, ranging from 0.17—0.20%. The steep initial gradient observed across all curves suggests the dominance

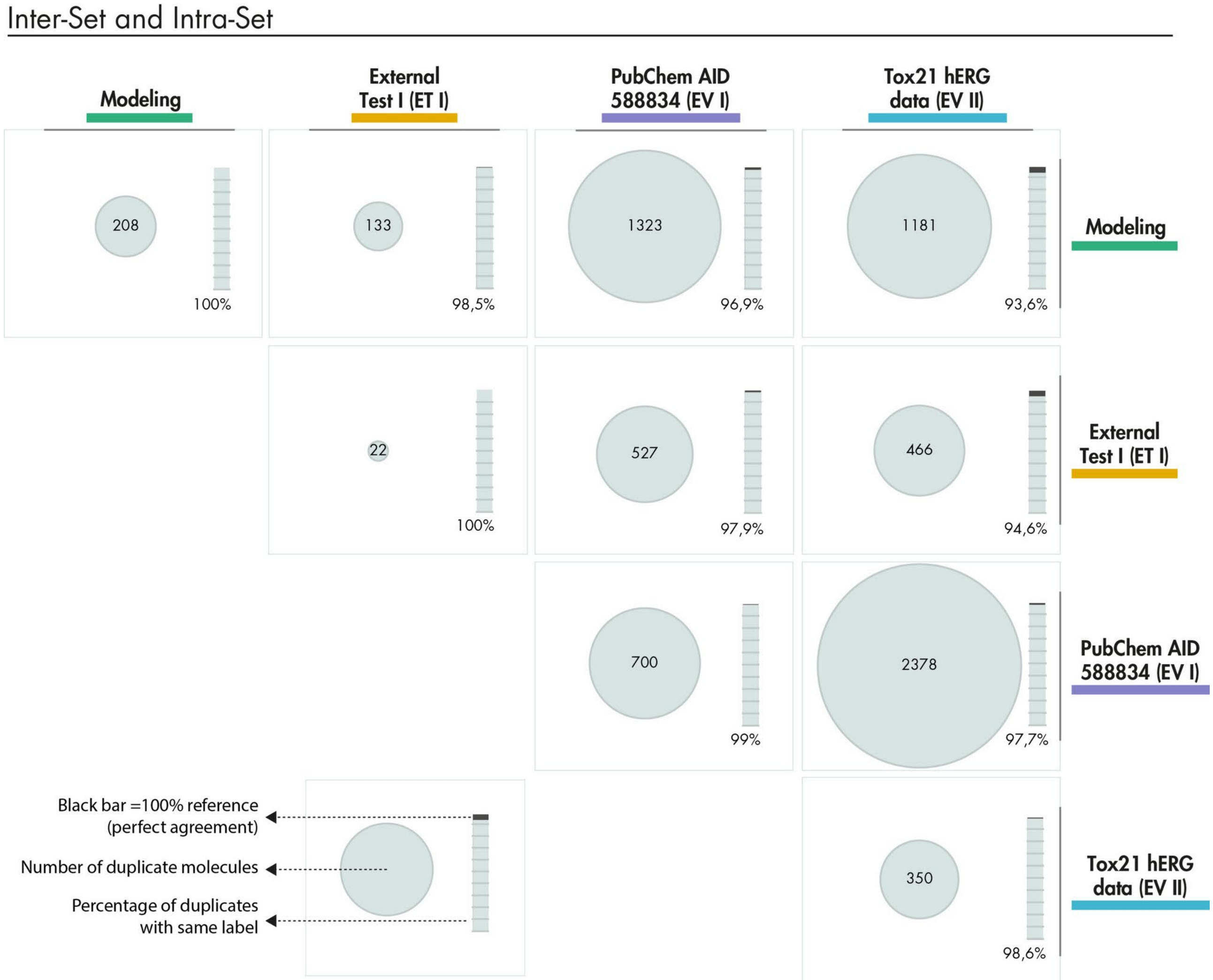


Fig. 2. Inter-set and intra-set hERG classification agreement for molecules with two or more experimental measurements. The diagonal of the matrix represents the classification agreement within each dataset (intra-set) after analyzing molecules with two or more experimental measurements (duplicates). The upper half of the matrix shows the number of overlapping molecules between sets (inter-set) and the percentage of duplicates with the same label for each set combination.

of specific structural motifs in the chemical space under study. The bottom diagram provides a matrix-like visualization of the number of shared Murcko scaffolds between datasets, where the size of each square is proportional to the number of common scaffolds. This visualization highlights scaffold diversity across datasets and their structural similarities, which are crucial for assessing the applicability and generalizability of predictive models.

Model performance and predictive power

The application of the methods described in the previous Variable Selection I subsection resulted in a refined list of 22 important variables (Selected Variables I). This reduced list was derived from an initial pool of 4,298 variables and became the initial pool of variables for the training of a new model in this lower dimensional space.

Utilizing an ensemble of 29 base classification models with these 22 variables demonstrated robust performance across all validation sets. Table 2 outlines the classification statistics for both internal set and ET I using the retrained model with the 22 variables. A detailed description of the model performance on Internal and ET I sets is provided in SI 4 and SI 5, respectively. The ACC remains consistently high, ranging from 0.87 to 0.90, and the model stands out for its balanced SP and SE values. To mitigate biased evaluation, statistics were also generated for 29 balanced internal and external test subsets. In each iteration, a balanced internal set was created by maintaining the active subset while changing the non-inhibitor samples. The results are presented in Table 2. In this experiment, the CK and MCC values exhibit a substantial increase, reaching up to 0.74. It is noteworthy that both metrics are influenced by the imbalance phenomenon.

Test Set	N	N variables	SE	SP	ACC	MCC	CK
Internal	20,343	22	0.83 ± 0.01	0.91 ± 0.01	0.90 ± 0.01	0.40 ± 0.01	0.32 ± 0.01
(Balanced)	1,345	22	0.83 ± 0.01	0.91 ± 0.01	0.87 ± 0.01	0.74 ± 0.01	0.73 ± 0.01
ET I	87,306	22	0.83 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.41 ± 0.01	0.33 ± 0.01
(Balanced)	5,976	22	0.83 ± 0.01	0.90 ± 0.01	0.87 ± 0.01	0.74 ± 0.01	0.74 ± 0.01

Table 2. Mean and standard deviation statistics for several metrics on internal, external and their respective balanced test sets, considering three independent runs of the ensemble with the majority vote as the output model. The abbreviations are detailed as follows: ET I: External Test Set (Ogura's Test Set)[13], N: number of samples, N variables: Number of input variables, SE: Sensitivity, SP: Specificity, ACC: Accuracy, MCC: Matthews correlation coefficient and CK: Cohen's kappa.

A comparative analysis of the results obtained in this work with other models available in the literature is presented in Table S3 in SI 1. This analysis includes models that: *i*) used the full hERG dataset[13,27], and *ii*) are commonly used and freely accessible such as DeepHIT[22] and the model of Delre et al.[14].

The ET I of the present study was used as the reference dataset for the comparison. Notable variations in performance metrics of models are found across different studies (Table S3 in SI 1). Ogura et al.[13] and Feng et al.[27] reported BACC values of 0.8 and 0.75 respectively, accompanied by a high SP of 0.99. However, their SE values were relatively low, standing at 0.67 and 0.51, respectively, indicating a suboptimal trade-off between SP and SE. Ogura et al. proposed an SVM model using 72 descriptors and ECFP_4 structural fingerprints. Feng et al. implemented two natural language processing methods, an autoencoder and a transformer, to embed molecular sequences. Both studies used the full hERG dataset from this study, originally compiled by Sato et al.[2].

The predictive power of both DeepHIT[22] and the model of Delre et al.[14] is moderate when applied to ET I. DeepHIT integrates three DL models with three different variables: molecular descriptors, fingerprints, and graph-based descriptors to produce high NPV. Delre et al.[14] used SMOTE, balanced random forest (BRF), gradient boosted trees (GBT) and SVM techniques. However, both studies rely on small, balanced datasets limiting the generalizability to highly imbalanced datasets like ET I. These results highlight the importance of the nature of training data in developing ML models for drug discovery. In the previously cited studies (DeepHit and Delre et al.), hERG data were collected from ChEMBL. Recent studies have shown that biological data in ChEMBL are biased toward the representation of active (inhibitor) compounds, whereas the opposite is observed in HTS data, which are generally highly imbalanced, with a small ratio of active to inactive compounds, as is the case in the present study[42]. The data distribution in this study is representative of a fully unbiased HTS approach.

The results of our study reveal a different performance profile. The model demonstrates a trade-off between SE (0.83) and SP (0.91), yet the challenge of class imbalance remains evident, as the MCC is around 0.4. This indicates that while the model is effective at identifying hERG inhibitors, the number of false positives (FP) remains high. This situation serves as a reason for the further implementation of the ISE map methodology (see below*)*, as a method to elucidate at which point the model preserves both SE and PPV for the inhibitor class. The balance between SE and SP in such scenarios is strongly dependent on the balancing method. For instance, in the paper by Ogura et al., a model-based approach was employed for handling class imbalance, specifically utilizing the class weight option during the training of the SVM. This method imposed a greater penalty for errors in the minority class, resulting in improved balance between SP and PPV. However, the SE remained low around 0.67[13].

The findings of the present study closely align with those reported by Ylipää et al.[43] in which, a graph neural network was applied as their top-performing model. By optimizing the discretization probability threshold through Youden's J-statistic, their model achieved more balanced performance metrics (ACC: 0.92, SE: 0.88, SP: 0.92). However, a direct comparison of the two models is not feasible due to differences in dataset composition. Specifically, they utilized only 93% of the Ogura dataset, excluding 5,638 chemicals during molecule preprocessing. Without access to detailed information regarding the excluded compounds, a meaningful comparison of model performance across the two studies remains unviable.

An essential consideration in anti-target prediction, such as hERG inhibition, is the cost associated with false predictions. For safety-focused in silico models, minimizing false negatives (FN) while maximizing true negatives is crucial to reduce cardiotoxicity risks during early drug development. The proposed model effectively reduces the likelihood of misclassifying a potentially hERG-inhibiting molecule as safe. As a result, the consensus methodology employed here enhances the reliability and accuracy of predictions for safety–critical targets, aligning with the stringent requirements of early-stage drug safety assessments. In summary, the comparison of our approach with other methods demonstrates that the model performs competitively. Our approach proves to be an effective and well-rounded model for the prediction of hERG inhibition, offering a solid balance between predictive performance and computational efficiency.

The OECD[40] emphasizes the importance of transparency and mechanistic relevance in computational models to facilitate interpretability. Our approach embodies these principles by operating in a low-dimensional, interpretable space that seamlessly balances mechanistic insights with robust predictive power. Specifically, the "Refinement: Prospective Validation and Model Interpretation" subsection presents a comprehensive variable importance analysis that highlights the key descriptors significantly contributing to the prediction of hERG activity. This analysis, conducted after applying two recursive variable selection methodologies, substantially enhances the interpretability of our model.

Enhancing predictive reliability of the XGBoost ensemble with the ISE map approach

The proposed model trains an ensemble of 29 base XGBoost classifiers and the ISE map provides a CL based on majority probability values and an ADL based on Euclidean distance for each sample. Figure 3 illustrates the distribution of key metrics, ACC, G-Mean, and MCC, across different CLs for both internal and external sets, based on partial and incremental statistical performance. Results reveal a clear trend: at the lowest consensus level (CL 0), where agreement among base models is minimal, statistical values for ACC, G-Mean, and MCC are at their lowest. In contrast, the highest CL (CL = 5), corresponds to the highest values for these metrics, indicating the most reliable performance. For instance, when Incremental Consensus at level 0 was evaluated on the external set, the ACC, G-Mean, and MCC were 90%, 87%, and 41%, respectively. These metrics improved when the same Consensus strategy was applied at level 5 (ACC = 98%, G-Mean = 94%, and MCC = 72%). This improvement was due to the ensemble making predictions in the chemical space where it is most accurate, resulting in a reduction of coverage from 100 to 61% and an increase in prediction reliability. The same tendency was observed with partial consensus (Fig. 3), but with worse performance results due to its restrictive nature. The partial and incremental CL results are presented in SI files 4 – 7.

The incremental CL approach offers a refined strategy for selecting compounds based on the statistical performance of the model. Specifically, when the objective is to identify compounds where the G-Mean of the model surpasses 90%, focusing on CL 4 and CL 5 proves effective. Compounds within these levels benefit from heightened predictive confidence, as they represent cases where nearly all base models consistently agree on class labels. At CL 4 and CL 5, approximately 78% of compounds across both internal and external sets meet this stringent reliability criterion, making these levels optimal for virtual screening aimed at prioritizing safe (hERG non-inhibitors) compounds, such as those in combinatorial or virtual chemical collections[44]. This high-confidence selection process ensures that the compounds retained for further testing exhibit strong predictive ACC, thus aligning well with early-stage safety assessments that demand high precision in anti-target screening.

The ISE-map organizes compounds on a discrete 2D grid based on their CL and ADL. The combination of the XGBoost ensemble with the ISE map improves previous approaches, because it incorporates, besides stratification by CL, AD information based on molecular similarity. According to the OECD requirements, computational models should have a well-defined AD to ensure their reliability.

Figure 4 provides a 2D representation of the ISE map, where the F-measure values for both the internal and ET I sets are plotted according to two key axes: CL on the x-axis and ADL on the y-axis. F-measure was selected because it provides a single score that balances the estimations of both PPV and SE in one number, making it particularly useful for imbalanced datasets. As CL increases, especially around levels 4 and 5 on the x-axis, and when ADL is 0 or 1, the color moves to darker shades of gray, indicating F-measure values greater than 0.55 and approaching 0.85. These values correspond to the best-classified compounds of the internal Set, whereas the worst-classified compounds are located at the opposite corner of the contour ISE map, with strongly decreasing coverage (see coverage values in Fig. 4). By analyzing the internal set ISE map, the strata with the best classification performance and highest coverage are (ADL, CL): (0, 4), (0, 5) and (1, 5). These strata can be selected and extrapolated to the ET I, from which the best classified molecules can then be sorted and triaged.

Figure 4 shows the resulting ET I ISE map with calculated F-measure cell values. Notably, there is a clear correspondence between the ISE map for the internal and ET I sets. Using the best-selected strata from the internal set, a coverage of 64% of the ET I was obtained (55,885 of 87,306 compounds), resulting in a global F-measure of 0.71, with ACC, CK, G-Mean, and MCC values of 0.98, 0.70, 0.93 and 0.72, respectively. These metrics represent a substantial improvement over our overall model (F-measure: 0.36, ACC: 0.90, CK: 0.33, G-Mean: 0.87, MCC: 0.41), increasing prediction reliability, despite the reduction in coverage from 100 to 64%.

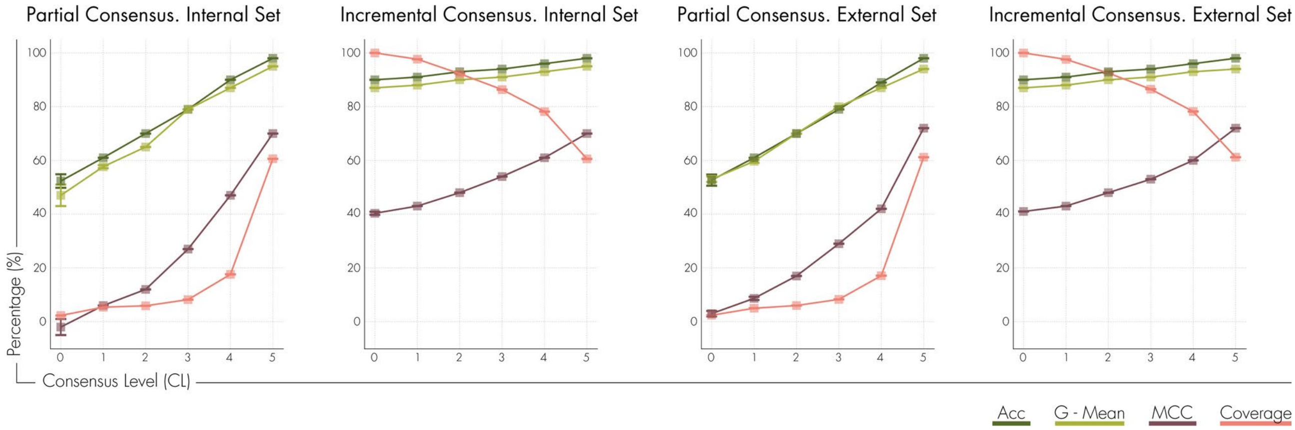
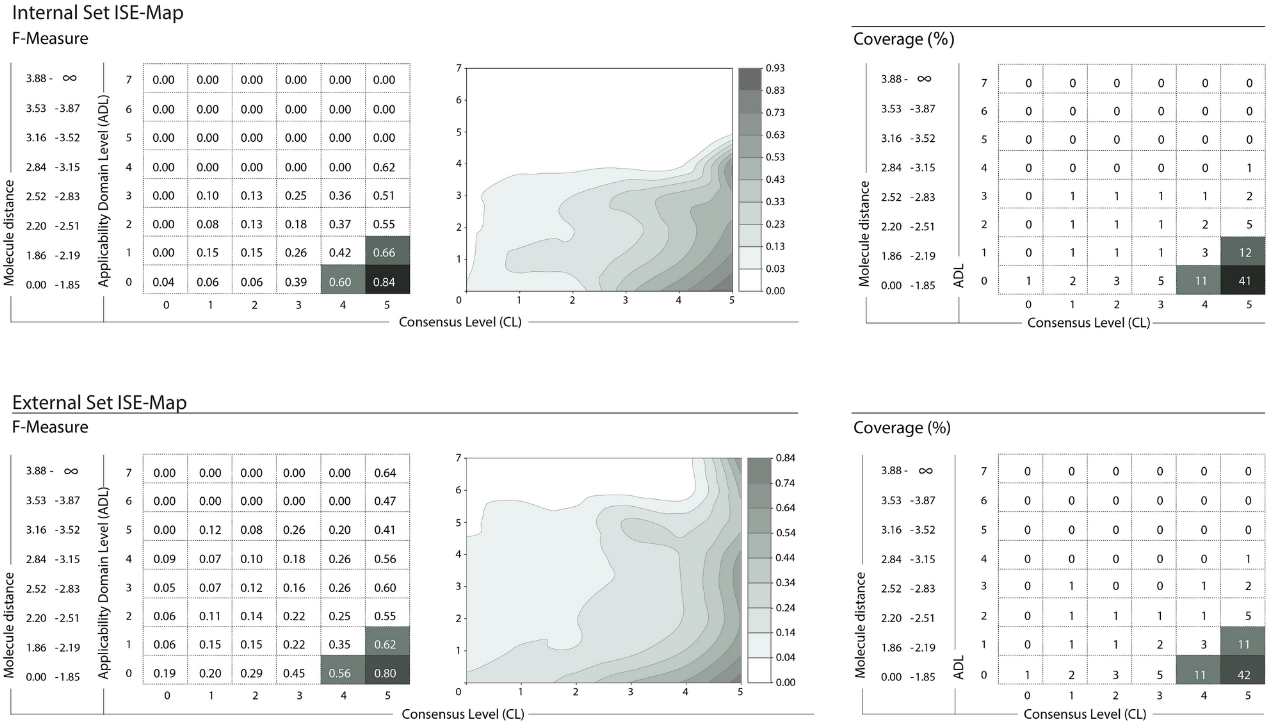


Fig. 3. Distribution of key metrics, Accuracy (ACC), Geometric Mean (G-Mean), Matthews Correlation Coefficient (MCC) and coverage, across different levels of consensus for both internal and external sets based on partial and incremental statistical performance. The error bars represent the standard deviation for each metric, calculated from three independent runs of each consensus model. All numerical statistics with means & standard deviations are available in SI files 4 and 5.



Internal Set ISE-Map

F-Measure

Molecule distance	Applicability Domain Level (ADL)	0	1	2	3	4	5
3.88 - ∞	7	0.00	0.00	0.00	0.00	0.00	0.00
3.53 -3.87	6	0.00	0.00	0.00	0.00	0.00	0.00
3.16 -3.52	5	0.00	0.00	0.00	0.00	0.00	0.00
2.84 -3.15	4	0.00	0.00	0.00	0.00	0.00	0.62
2.52 -2.83	3	0.00	0.10	0.13	0.25	0.36	0.51
2.20 -2.51	2	0.00	0.08	0.13	0.18	0.37	0.55
1.86 -2.19	1	0.00	0.15	0.15	0.26	0.42	0.66
0.00 -1.85	0	0.04	0.06	0.06	0.39	0.60	0.84

Consensus Level (CL)

Coverage (%)

Molecule distance	ADL	0	1	2	3	4	5
3.88 - ∞	7	0	0	0	0	0	0
3.53 -3.87	6	0	0	0	0	0	0
3.16 -3.52	5	0	0	0	0	0	0
2.84 -3.15	4	0	0	0	0	0	1
2.52 -2.83	3	0	1	1	1	1	2
2.20 -2.51	2	0	1	1	1	2	5
1.86 -2.19	1	0	1	1	1	3	12
0.00 -1.85	0	1	2	3	5	11	41

Consensus Level (CL)

External Set ISE-Map

F-Measure

Molecule distance	Applicability Domain Level (ADL)	0	1	2	3	4	5
3.88 - ∞	7	0.00	0.00	0.00	0.00	0.00	0.64
3.53 -3.87	6	0.00	0.00	0.00	0.00	0.00	0.47
3.16 -3.52	5	0.00	0.12	0.08	0.26	0.20	0.41
2.84 -3.15	4	0.09	0.07	0.10	0.18	0.26	0.56
2.52 -2.83	3	0.05	0.07	0.12	0.16	0.26	0.60
2.20 -2.51	2	0.06	0.11	0.14	0.22	0.25	0.55
1.86 -2.19	1	0.06	0.15	0.15	0.22	0.35	0.62
0.00 -1.85	0	0.19	0.20	0.29	0.45	0.56	0.80

Consensus Level (CL)

Coverage (%)

Molecule distance	ADL	0	1	2	3	4	5
3.88 - ∞	7	0	0	0	0	0	0
3.53 -3.87	6	0	0	0	0	0	0
3.16 -3.52	5	0	0	0	0	0	0
2.84 -3.15	4	0	0	0	0	0	1
2.52 -2.83	3	0	1	0	0	1	2
2.20 -2.51	2	0	1	1	1	1	5
1.86 -2.19	1	0	1	1	2	3	11
0.00 -1.85	0	1	2	3	5	11	42

Consensus Level (CL)

Fig. 4. ISE map of F-measure values per stratum represented as a 2D contour plot for the **(a)** Internal Set, and **(b)** External Test Set (ET I). In the contour plot, the color gradient ranges from lighter shades of gray (low F-measure values) to darker shades of gray (high F-measure values). The contour lines delineate these transitions between the F-measure values throughout the plot. In the table, F-measures > 0.55 and coverage > 10% are marked with a gray background. The abbreviations are detailed as follows: ADL: Applicability Domain Level and ISE: Isometric Stratified Ensemble.

By organizing data into strata based on CLs and incorporating Nearest Neighbor distance values, the ISE map not only establishes the model AD but also facilitates a nuanced understanding of prediction confidence, which is another standout feature of this study. This approach contributes to the broader field of predictive toxicology modeling by providing a graphical tool for assessing the reliability of model predictions across different consensus and AD levels.

Overall, our methodology effectively addresses data imbalance and enhances model reliability through a refined validation framework incorporating the ISE map. While similar performance levels have been reported in other studies, our approach reduces FN, critical in early drug discovery, and provides a robust method for hERG screening. The refined selection method based on the ISE map helps identify promising molecules while minimizing the risk of selecting those with undesired hERG inhibition.

Refinement: Prospective validation and model interpretation

Reducing the number of input variables lowers computational costs and improves interpretability. In this study, 22 descriptors were selected from an initial set of 4,298 using the RVS technique, followed by RVE to identify the minimum number of variables needed for a compact, interpretable model of hERG inhibition.

The statistical analysis described in Variable Selection II in Method section indicated that only eight variables are strictly significant for predicting the target variable using the current methodology. A graphical representation is provided in Figure S2 in SI 1. The sorted list of the most important variables for hERG prediction is shown in Table S4 in SI 1. Additionally, Figure S3 in SI 1 presents the Pearson and Spearman rank correlation coefficients for each pair of variables, illustrating low correlations between them. These results suggest that the two variable selection algorithms effectively reduce model complexity by selecting the most important and non-redundant variables, thereby enhancing interpretability.

Table 3 shows performance comparisons when reducing descriptors from 4,298 ((RVS + XGB)_a) to 22 ((RVS + XGB)_b), and ultimately to 8 ((RVS + XGB)_c), highlighting the interpretability gains with fewer variables. The model developed from the complete pool of descriptors exhibited only a marginal improvement over the model with 22 descriptors across the three external sets, demonstrating their equivalence in performance. However, the second model was adopted and further developed, adhering to the principle of parsimony and favoring a low-dimensional descriptor space. Table 3 also summarizes the results obtained from external sets using these model predictions against the output of the ISE map methodology, based on the previously selected strata. Our XGBoost ISE approach outperforms the overall models, enhancing screening power despite a decrease in coverage.

Set	Method	Input Variables	Cov (%)	IR	SE	SP	ACC	MCC	F-1	CK
ET	(RVS + XGB)_a	4,298	100	28	0.85	0.92	0.92	0.46	0.42	0.39
	(RVS + XGB)_b	22	100	28	0.83	0.90	0.90	0.41	0.36	0.33
	(RVS + XGB)_c	8	100	28	0.80	0.85	0.85	0.32	0.27	0.22
	(RVS + XGB)_b + ISE	22	64	36	0.88	0.98	0.98	0.72	0.71	0.70
EV I	(RVS + XGB)_a	4,298	100	11	0.75	0.96	0.94	0.65	0.68	0.64
	(RVS + XGB)_b	22	100	11	0.74	0.94	0.93	0.60	0.63	0.59
	(RVS + XGB)_c	8	100	11	0.71	0.93	0.91	0.53	0.56	0.51
	(RVS + XGB)_b + ISE	22	69	12	0.82	0.98	0.97	0.80	0.81	0.80
EV II	(RVS + XGB)_a	4,298	100	24	0.84	0.95	0.94	0.55	0.54	0.51
	(RVS + XGB)_b	22	100	24	0.77	0.94	0.94	0.50	0.49	0.46
	(RVS + XGB)_c	8	100	24	0.75	0.93	0.92	0.44	0.42	0.39
	(RVS + XGB)_b + ISE	22	71	28	0.97	0.97	0.97	0.67	0.67	0.65

Table 3. XGBoost (XGB) ensemble performance on external sets following the application of variable selection procedures (RVS). (RVS + XGB)_a is the model obtained using all initial molecular descriptors (4,298 variables) with RVS, (RVS + XGB)_b is the model obtained using a reduced initial set of molecular descriptors (22 variables), (RVS + XGB)_c is the model with the lowest dimensionality (8 variables) and (RVS + XGB) _b + ISE is the model derived from a reduced initial set of molecular descriptors (22 variables) with RVS, considering only the predictions made in the previously selected strata (ADL, CL): (0, 4), (0, 5), and (1, 5). Input variables are the initial pool of molecular descriptors from which the model is built. The abbreviations are detailed as follows: Cov: Coverage, IR: Imbalance Ratio, SE: Sensitivity, SP: Specificity, ACC: Accuracy, MCC: Matthews correlation coefficient, F-1: F-measure and CK: Cohen's kappa. ET I is the External Test Set (Ogura's Test Set)[13], while EV I and EV II are external validation sets extracted from PubChem assay AID 588334[36] and Tox21 hERG data[36], respectively.

Rank	Variables	Median Value		Type	Description
		Inhibitor	Non-inhibitor		
1	peoe_VSA8	30.714	17.92	VSA Descriptors	P_VSA- like on PEOE (partial charges), bin 8
2	ESOL	-4.672	-3.829	Molecular Properties	Estimated Solubility (logS) for aqueous solubility using LOGPcons
3	SdssC	0	-0.217	Atom-type E-state indices	Sum of E-states of atoms of type = C < within a molecule
4	MaxssO	4.880	4.780	Atom-type E-state indices	Maximum atom-type E-State: -O-
5	nRNR2	1 (0–4)	0 (0–4)	Functional group counts	Number of tertiary amines (aliphatic)
6	MATS1i	-0.139	-0.108	2D autocorrelations	Moran autocorrelation of lag 1 weighted by ionization potential
7	nRNHR	0 (0–2)	0 (0–4)	Functional group counts	Number of secondary amines (aliphatic)
8	nRNH2	0 (0–2)	0 (0–5)	Functional group counts	Number of primary amines (aliphatic)

Table 4. List of selected variables/descriptors from the Recursive Variable Elimination algorithm, including median descriptor values for hERG inhibitors and non-inhibitors in the training set. A Mann–Whitney U test showed significant differences ($p < 0.05$) in continuous descriptors between hERG inhibitors and non-inhibitors. A Chi-squared analysis revealed a significant association between the variables nRNH2, nRNHR and nRNR2 and the hERG activity ($p < 0.001$). The rank refers to the order of importance of the 8 descriptors included in the lowest dimensionality model. All descriptors shown here were calculated with alvaDesc software, except for peoe_VSA8, which was calculated using RDKit software. In parentheses, the maximum and minimum values of the molecular descriptors nRNH2, nRNHR, and nRNR2 are noted.

Table 3 also presents a model with significantly lower dimensionality (8 variables) that can predict outcomes for EV I and EV II with strong statistical performance. These results support the ability of the model to predict hERG inhibition potential, as EV I and EV II were derived from standardized assays. The predictive power of the model remained robust despite the reduced set of descriptors, emphasizing the efficiency of the employed methodology. However, the model with 22 variables remains the one that achieves the best balance between the number of variables used and model performance across all external sets.

To gain insights into the structure-hERG inhibition relationship, the 8 most important variables were examined. Table 4 lists these descriptors, belonging to five logical blocks: molecular properties, autocorrelation descriptors, atom-type electrotopological state (E-state) indices, VSA descriptors and functional group counts. Although no single descriptor explains hERG activity alone, their combination provides valuable insights.

The variable MATS1i is a Moran (M) autocorrelation descriptor with lag 1 that reflects spatial autocorrelation in a molecular graph where atoms are weighted by ionization potential (i)[45]. When lag = 1, MATS provides a measure of Nearest Neighbor correlations; when lag = 2, next Nearest Neighbor, and so on. In general,

-1 < MATS < 1. Positive autocorrelations correspond to positive values of the Moran coefficient, indicating that atoms with high ionization potential are likely to follow each other along the atom sequence at topological distance of 1. Conversely, negative autocorrelation produces negative values, meaning that a high value is followed by a low value, or vice versa, at lag = 1. For instance, transformations that incorporate atoms with high ionization potential, such as nitrogen and/or oxygen, may reduce hERG inhibitory activity when these atoms are adjacent in the molecular structure. These changes are reflected in the MATS1i descriptor, which shows higher median values in non-inhibitors than in inhibitors.

Other significant descriptors include MaxssO, the maximum E-state value among all -O- groups in the molecule, and SdssC, the sum of E-state values for carbon atoms with sp2 hybridization. E-state values capture both electronic and topological information, reflecting the intrinsic electronic state of the carbon and oxygen atom and the influence of surrounding atoms. The E-state of an atom S_i is formulated as the sum of its intrinsic state I_i, and a perturbation term ΔI_{ij} which represents the influence of the j-th atom on the i-th intrinsic state. This perturbation term acts as an electronegative gradient, with its sign indicating the direction of influence from surrounding atom intrinsic states. Thus, large positive E-state values are associated with highly electronegative and/or terminal atoms that lie on the periphery of the molecule. In contrast, small or negative E-state values correspond to atoms involved in sigma bonding, atoms that are located in the core of the molecule or atoms close to highly electronegative atoms[45]. Table 4 shows that hERG non-inhibitors exhibit lower median values for both descriptors compared to the inhibitors. For instance, introducing = C < groups near highly electronegative atoms and/or in the interior of the molecule reduces these descriptor values, thereby contributing to non-inhibition. This reduces the electronic accessibility of these atoms and limits their interaction with other molecules, such as the hERG channel. These findings align with previous studies that emphasize the role of electrostatic interactions with amino acid residues around the hERG cavity, such as Y652, as the key mechanism in hERG inhibition[46].

The peoe_VSA8 descriptor exhibits a similar trend to that of E-state indices. This descriptor has higher median values for inhibitor compounds compared to non-inhibitors. This descriptor is calculated as the sum of atomic contributions to the van der Waals surface area (VSA) from atoms within specific ranges, in this case, bin 8[47]. This suggests that the contact surfaces related to charge localization are significant factors in hERG inhibition. PEOE charges depend on the connectivity of the input structures, including elements, formal charges, and bond orders. The observation that uncharged molecules exhibit suboptimal hERG inhibition may be related to their electrostatic characteristics. Such molecules may lack the essential charge distribution patterns required for effective hERG inhibition.

The ESOL descriptor is primarily designed to predict the aqueous solubility of compounds based on their molecular structure, utilizing properties such as the octanol/water logP, molecular weight, etc.[48]. The trend observed in the ESOL descriptor shows that many hERG inhibitors tend to exhibit lower aqueous solubility. This finding is also related to the behavior of the nRNR2 descriptor, which is directly associated with the presence of tertiary amines in the structural moieties. The presence of basic nitrogen atoms with additional hydrophobic groups is a classical characteristic of many hERG blockers. This is justified by the hydrogen bond between the protonated nitrogen of the channel blocker and the carbonyl oxygen of hERG residue T623[46].

Despite the non-linear relationships between the selected variables and hERG inhibition, certain tendencies in the distribution of descriptors can still be observed, as described above. However, their practical implications necessitate further analysis.

Conclusions

This study presents a robust and comprehensive framework for the QSAR modeling of hERG inhibition. By generating multiple balanced training subsets, we ensured that our models were not skewed toward the majority class, thus enhancing their predictive performance, particularly reducing the false negative rate and minimizing potential cardiotoxicity risks in early drug discovery. The use of recursive methods for variable selection enhances model interpretability and is important for translating computational insights into actionable knowledge in drug design. Prospective validation on two external datasets offers a comprehensive assessment of the model generalization capabilities, showing robust predictive power even with fewer descriptors. A standout feature of this research is the application of the ISE map for estimating the applicability domain and prediction confidence. By organizing data into strata based on consensus level and incorporating Nearest Neighbor distances, the ISE map serves as a model-agnostic method that enhances model interpretability and facilitates a nuanced understanding of prediction confidence. This methodological innovation contributes significantly to the field of predictive modeling by providing a graphical tool for assessing the reliability of model prediction across various consensus levels. The XGB + ISE map strategy provides an effective approach to identify and prioritize promising molecules in drug discovery campaigns with reduced hERG inhibition risk.

Data availability

All data generated or analyzed during this study are included in the main text of this article and its Supplementary Information files.

Code availability

The computational method used to generate the results in this study is accessible through a web server deployed at https://www.hypercubane.re/. This server allows users to submit data and receive predictions based on the hERG model described in this manuscript. Detailed instructions for using the server are available on the web page. The code is available under license for academic research and for commercial use subject to negotiation. For further information, please contact the corresponding author CM.

Received: 20 December 2024; Accepted: 21 April 2025
Published online: 04 May 2025

References

1. Creanza, T. M. et al. Structure-based prediction of hERG-related cardiotoxicity: A benchmark study. *J. Chem. Inf. Model.* **61**, 4758–4770 (2021).
2. Sato, T., Yuki, H., Ogura, K. & Honma, T. Construction of an integrated database for hERG blocking small molecules. *PLoS ONE* **13**, e0199348 (2018).
3. Mamoshina, P., Rodriguez, B. & Bueno-Orovio, A. Toward a broader view of mechanisms of drug cardiotoxicity. *Cell Rep. Med.* **2**, 100216 (2021).
4. Kalyaanamoorthy, S. & Barakat, K. H. Development of safe drugs: The hERG challenge. *Med. Res. Rev.* **38**, 525–555 (2018).
5. Kalyaanamoorthy, S. et al. A structure-based computational workflow to predict liability and binding modes of small molecules to hERG. *Sci. Rep.* **10**, 16262 (2020).
6. Cai, C. et al. Deep learning-based prediction of drug-induced cardiotoxicity. *J. Chem. Inf. Model.* **59**, 1073–1084 (2019).
7. Tran, T. T. V., Surya Wibowo, A., Tayara, H. & Chong, K. T. Artificial intelligence in drug toxicity prediction: Recent advances, challenges, and future perspectives. *J. Chem. Inf. Model.* **63**(2628), 2643 (2023).
8. Kang, M. et al. A Review of the ethical use of animals in functional experimental research in china based on the 'four r' principles of reduction, replacement, refinement, and responsibility. *Med. Sci. Monit.* **28**, e938807 (2022).
9. Lee, K. H., Lee, D. W. & Kang, B. C. The 'R' principles in laboratory animal experiments. *Lab. Anim. Res.* **36**, 45 (2020).
10. Villoutreix, B. O. & Taboureau, O. Computational investigations of hERG channel blockers: New insights and current predictive models. *Adv. Drug. Deliv. Rev.* **86**, 72–82 (2015).
11. Colatsky, T. et al. The Comprehensive in Vitro Proarrhythmia Assay (CiPA) initiative - Update on progress.*J. Pharmacol. Toxicol. Methods***81**, 15–20 (2016).
12. Arab, I. & Barakat, K. ToxTree: descriptor-based machine learning models for both hERG and Nav1.5 cardiotoxicity liability predictions. *arXiv preprint* arXiv:2112.13467 (2021)
13. Ogura, K., Sato, T., Yuki, H. & Honma, T. Support vector machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II. *Sci. Rep.* **9**, 12220 (2019).
14. Delre, P. et al. Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. *Front. Pharmacol.* **13**, 951083 (2022).
15. Hsiao, Y., Su, B. H. & Tseng, Y. J. Current development of integrated web servers for preclinical safety and pharmacokinetics assessments in drug development. *Brief. Bioinform.* **22**(3), bbaa160 (2021).
16. Konda, L. S. K., Keerthi Praba, S. & Kristam, R. hERG liability classification models using machine learning techniques. *Comput. Toxicol.* **12**, 100089 (2019).
17. Siramshetty, V. B. et al. Critical assessment of artificial intelligence methods for prediction of hERG channel inhibition in the "Big Data" era. *J. Chem. Inf. Model.* **60**, 6007–6019 (2020).
18. Tian, S., Li, Y., Wang, J., Zhang, J. & Hou, T. ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Mol. Pharm.* **8**, 841–851s (2011).
19. Shan, M. et al. Predicting hERG channel blockers with directed message passing neural networks. *RSC Adv.* **12**, 3423–3430 (2022).
20. 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.* **13**, 60 (2021).
21. Zhang, X., Mao, J., Wei, M., Qi, Y. & Zhang, J. Z. H. HergSPred: Accurate classification of hERG blockers/nonblockers with machine-learning models. *J. Chem. Inf. Model.* **62**, 1830–1839 (2022).
22. Ryu, J. Y., Lee, M. Y., Lee, J. H., Lee, B. H. & Oh, K.-S. DeepHIT: A deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics* **36**, 3049–3055 (2020).
23. Sun, H. et al. Highly predictive and interpretable models for PAMPA permeability. *Bioorg. Med. Chem.* **25**, 1266–1276 (2017).
24. Xiong, G. et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **49**, W5–W14 (2021).
25. Venkatraman, V. FP-ADMET: A compendium of fingerprint-based ADMET prediction models. *J. Cheminform.* **13**, 75 (2021).
26. Ishihara, K., Sone, H. & Hayamizu, K. Gradient boosting machine-based model for predicting hERG K+ channel inhibitory activities. *J. Comp. Sci. Appl. Inform. Technol.* **5**, 1–9 (2022).
27. Feng, H. & Wei, G.-W. Virtual screening of drugbank database for hERG blockers using topological laplacian-assisted AI models. *Comput. Biol. Med.* **153**, 106491 (2023).
28. Babajide Mustapha, I. & Saeed, F. Bioactive molecule prediction using extreme gradient boosting. *Molecules* **21**, 983 (2016).
29. Casanova-Alvarez, O., Morales-Helguera, A., Cabrera-Pérez, M. Á., Molina-Ruiz, R. & Molina, C. A novel automated framework for QSAR modeling of highly imbalanced leishmania high-throughput screening data. *J. Chem. Inf. Model.* **61**, 3213–3231 (2021).
30. Molina, C., Ait-Ouarab, L. & Minoux, H. Isometric stratified ensembles: a partial and incremental adaptive applicability domain and consensus-based classification strategy for highly imbalanced data sets with application to colloidal aggregation. *J. Chem. Inf. Model.* **62**, 1849–1862 (2022).
31. Falcón-Cano, G., Molina, C. & Cabrera-Pérez, M. Á. Reliable prediction of caco-2 permeability by supervised recursive machine learning approaches. *Pharmaceutics* **14**, 1998 (2022).
32. Kamel, M. *et al.* Automated workflows for data curation and standardization of chemical structures for QSAR modeling. In (American Chemical Society meeting, Orleans, LA, 2018). https://doi.org/10.13140/RG.2.2.10944.84484. 2018
33. RDKit Nodes for KNIME (trusted extension) KNIME. https://www.knime.com/rdkit.
34. Mauri, A. & Bertola, M. Alvascience: A New software suite for the QSAR workflow applied to the blood-brain barrier permeability. *Int. J. Mol. Sci.* **23**, 12882 (2022).
35. alvaDesc - KNIME - Alvascience. https://www.alvascience.com/knime-alvadesc/.
36. National Center for Biotechnology Information. PubChem Bioassay Record for AID 588834, qHTS Assay for Small Molecule Inhibitors of the Human hERG Channel Activity. https://pubchem.ncbi.nlm.nih.gov/bioassay/588834 (2017).
37. NCATS, EPA, NIEHS, NTP, & FDA,. Tox21 Data Challenge. Human Ether-a-go-go-Related Gene (hERG) channel small molecule antagonists: fluorescence cell-based qHTS assay U.S. Tox21 Program. https://tripod.nih.gov/pubdata/.
38. Langdon, S. R., Brown, N. & Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* **51**, 2174–2185 (2011).
39. XGBoost Documentation — xgboost 2.1.1 documentation. https://xgboost.readthedocs.io/en/stable/.
40. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. *OECD Series on Testing and Assessment. No 69. OECD Publishing. Paris* (2014).
41. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. *Mol. Framew. J. Med. Chem.* **39**, 2887–2893 (1996).
42. Zakharov, A. V., Peach, M. L., Sitzmann, M. & Nicklaus, M. C. QSAR modeling of imbalanced high-throughput screening data in PubChem. *J. Chem. Inf. Model* **54**, 705–712 (2014).
43. Ylipää, E. et al. hERG-toxicity prediction using traditional machine learning and advanced deep learning techniques. *Curr. Res. Toxicol.* **5**, 100121 (2023).
44. Compound Collections - Enamine. https://enamine.net/compound-collections.

45. Todeschini, R. & Consonni, V. *Handbook of molecular descriptors* (John Wiley & Sons Ltd, 2000). https://doi.org/10.1002/9783527613106.ch1a.
46. Garrido, A., Lepailleur, A., Mignani, S. M., Dallemagne, P. & Rochais, C. hERG toxicity assessment: Useful guidelines for drug design. *Eur. J. Med. Chem.* **195**, 112290 (2020).
47. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **18**, 464–477 (2000).
48. Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).

Acknowledgements

All figures presented in the main text and Supplementary Information were originally created by the authors, G.F.-C., A.M.-H., and C.M. We thank Silvia Romanelli for her professional contribution to the graphical design of this work. The authors also acknowledge KNIME and its contributors for making the KNIME data mining environment available free of charge, as well as Alvascience for providing an academic license for alvaDesc.

Author contributions

Research design: C.M., G.F-C. & A.M-H.; Methodology: C.M., G.F-C., A.M-H. & H.L.; Data Curation: H.L. & G.F-C.; Data Analysis: C.M., G.F-C. & A.M-H.; Manuscript Drafting: G.F-C., A.M-H, C.M. & M-A.C-P. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare the following conflict of interest: Pikaïros is a commercial entity, and the corresponding author, CM, is the sole shareholder of the company. The other authors declare no competing financial interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-99766-3.

Correspondence and requests for materials should be addressed to C.M.