Scientific Reports

https://doi.org/10.1038/s41598-026-39771-2

Article in Press

Interpretable machine learning rationalizes carbonic anhydrase inhibition via conformal and counterfactual prediction

Received: 2 November 2025

Accepted: 6 February 2026

Published online: 11 February 2026

Cite this article as: Ghamsary M.S., Rayka M. & Naghavi S.S. Interpretable machine learning rationalizes carbonic anhydrase inhibition via conformal and counterfactual prediction. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39771-2

Masoumeh Shams Ghamsary, Milad Rayka & S. Shahab Naghavi

Interpretable Machine Learning Rationalizes Carbonic Anhydrase Inhibition via Conformal and Counterfactual Prediction

Masoumeh Shams Ghamsary [These authors contributed equally][1], Milad Rayka [These authors contributed equally][1], and S. Shahab Naghavi [Corresponding Author: s_naghavi@sbu.ac.ir][1]

[1]Department of Physical and Computational Chemistry, Shahid Beheshti University, Tehran 1983969411, Iran

ABSTRACT

Human carbonic anhydrase (hCA) isoforms IX and XII are promising anticancer targets. Yet, their selective inhibition remains elusive due to close similarity with the abundant hCA II, whose off-target inhibition causes harmful side effects. Here, we introduce an interpretable machine learning framework to predict inhibition across hCA II, IX, and XII. To address this issue, our approach combines rigorous data curation, systematic benchmarking of classical and deep learning models, and integration of conformal prediction for uncertainty quantification with counterfactual explanations for molecular interpretability. After extensive benchmarking, we find that Support Vector Machines with extended-connectivity fingerprints consistently outperform more complex models, underscoring the importance of data quality and validation over algorithmic complexity. Here, conformal prediction provides rigorous activity estimation, while counterfactual analysis rationalizes structural features governing isoform selectivity, together enabling interpretable guidance for inhibitor design. To further test our model capability, we examine it on SLC-0111, as a selective inhibitor, which leads to a compatible result with the experiment. Our model reiterates experimental findings that modifications in the tail region strongly affect molecular selectivity, emphasizing the tail group as a key structural determinant for differentiating inhibitor activity among hCA isoforms II, IX, and XII. To facilitate adoption, we also release `CAInsight`, a user-friendly software with a graphical interface for virtual screening and generative design of a selective hCA inhibition.

Keywords: Human Carbonic Anhydrase, Machine Learning, Deep Learning, Selective Inhibitor, Activity, Interpretability, Uncertainty Quantification

Introduction

Target-specific drug delivery poses a central challenge in modern therapeutics, where enhancing binding affinity and selectivity toward disease-related targets is essential to minimize off-target effects[1]. Precise target binding is vital, especially in life-threatening diseases, where unintended drug effects overshadow the disease itself[2]. Cancer is one such disease: a leading cause of death, taking ten million lives annually, according to the World Health Organization (WHO)[3]. Cancer becomes lethal at the hypoxic stage, when rapid tumor growth drains oxygen from surrounding tissue[4]. Recent research has identified strategies to counter hypoxia, including inhibition of the human carbonic anhydrase (hCA) enzyme, the enzyme that suppresses tumor growth and enhances treatment efficacy[5–8]. These findings establish the hCA enzyme as a key target for next wave of cancer therapies[9].

Among the 16 isoforms of hCAs, hCA IX and XII help cells survive in hypoxic conditions[10]. But, isoform II has active-site similar to IX and XII and is widespread in many cell types, making it nearly impossible to target IX and XII without also inhibiting II[11]. Such off-target binding to isoform II triggers side effects including metabolic acidosis, electrolyte imbalance, and vision disturbances[12]. Here, lab research and computational methods help target-specific drug design yet face two hurdles[13]: (i) the large molecular structure of enzymes complicates accurate modeling, and (ii) the extensive chemical space of enzyme-ligand interactions makes experimental screening impossible.

Given these hurdles, data-driven algorithms offer a powerful route forward. Decades of biological data, combined with advances in computing power and algorithm design, have produced robust machine learning (ML) and deep learning (DL) techniques[14–16]—an achievement recognized by the 2024 Nobel Prize in Chemistry for breakthroughs in protein design[17].

ML and DL have also revolutionized drug-target prediction over the past decade[18–20]. Yet, despite their widespread use in target-specific drug design[21–23], work on the hCA topic remain limited to a mere two studies. In the first, Galati et al.[24] used bioactivity data from the PubChem Bioassay database[25] to train machine-learning classification models that distinguish isoforms

IX and II. They represented each molecule as Extended-Connectivity Fingerprints (ECFPs)[26] and applied only Random Forest (RF)[27] and Support Vector Machine (SVM)[28] algorithms. In the second study, Tinivella et al.[29] trained a range of traditional ML algorithms (e.g., Logistic Regression (LR)[28], SVM, and RF) using ChEMBL bioactivity data to classify molecules binding to isoforms II, IX, and XII. Their models relied solely on RDKit molecular descriptors for feature generation.

The literature shows recurring flaws: poor data preprocessing, limitted use of feature vectors and algorithms, neglect of data imbalance, weak hyperparameter tuning and validation, and little attention to model uncertainty or interpretability. This study presents a rigorous pipeline (Fig. 1) to close these gaps. In this work, our contributions are:

- Development of three ligand-based binary classification models to predict inhibition of hCA isoforms II, IX, and XII.
- Implementation of a comprehensive data preprocessing pipeline, including sanitization, standardization, deduplication, and data imbalance handling.
- Systematic benchmarking of diverse machine and deep learning models—from Logistic Regression to Graph Neural Networks—across multiple molecular representations.
- Extensive model evaluation, including hyperparameter optimization, validation procedures, and comparison of random versus scaffold-based data splitting strategies to assess robustness and generalizability.
- Integration of conformal prediction for uncertainty quantification and counterfactual explanations for model explainability.
- Release of `CAInsight`, a user-friendly software for exploring the developed tools and facilitating virtual screening workflows.

The results evidently demonstrate that the SVM model using ECFP, enhanced by conformal prediction and counterfactual explanation, predicts compound inhibition across all three isoforms. All code to reproduce this study and the `CAInsight` software are available at https://github.com/miladrayka/hca_ml.

Methods

Database Retrieval

We retrieve hCA II, hCA IX, and hCA XII bioactivity datasets from ChEMBL (Sept 2024) using its official Python client (v0.10.9) (Fig. 1a)[30]. The dataset contains compounds with measured binding affinities for hCA II, hCA IX, and hCA XII with UniProt IDs “P00918”, “Q16790”, and “O43570”, respectively.

Binding affinities are expressed as K_i, K_d, and IC50 values: K_i is the inhibition constant, K_d the dissociation constant, and IC50 the concentration required to inhibit a biological process by 50%. To maximize dataset size and chemical diversity, K_i, K_d, and IC_{50} values are aggregated. This is a commonly used approach in ligand or structure-based modeling to ensure a broad representation of structural scaffolds and mitigate the data deficiency[31–33]. While these metrics differ experimentally, the logarithmic transformation helps place the values on a comparable scale. By using a binarization threshold (see Dataset Split subsection), we minimize potential discripency, as the model focuses on distinguishing broad activity classes rather than exact numerical differences.

Entries with non-molar activity measurements, such as percent inhibition or fold change, are excluded, and all units are limited to nanomolar (nM). Only records with activity relationships marked “=” (verified) or “>” (inactive) are kept; those with missing values are removed. Activity values are given as negative logarithms ($pK_{i/d/IC50}$).

SMILES are processed using sanitization and standardization functions of the `Datamol` package (v0.12.5)[34], ensuring chemically valid molecules with canonical representations. Then, the “Structure Filter” utility in `RDkit` (v2023.9.5)[35] is then used to retain molecules with a primary sulfonamide zinc-binding group (ZBG), removing allosteric inhibitors and uncommon ZBG molecules[29].

Dataset Split

Following Tinivella et al.[29], molecules are labeled “active” when their activity is < 20 nM and “inactive” when it is > 100 nM (Fig. 1b). These stricter thresholds, compared with more traditional cutoffs, align with the potency range typically pursued for hCA inhibitors[36]. Compounds in the intermediate range (20–100 nM) are excluded because values near the decision boundary introduce label noise that can degrade classifier performance[37]. Excluding such borderline cases yields cleaner class labels and structure–activity relationships. As these thresholds lead to class imbalance (also reported by Tinivella et al.[29]), we apply random oversampling during training to prevent bias toward the majority class. The dataset is split using random or scaffold-based methods. Random splitting can inflate performance estimates due to structural similarity[38]. To prevent this, scaffold-based splitting groups molecules by their Bemis-Murcko scaffolds[39] using `DeepChem` (v2.7.1)[40]. This method enforces structural separation across training, validation, and test sets while preserving scaffold diversity and reducing bias.

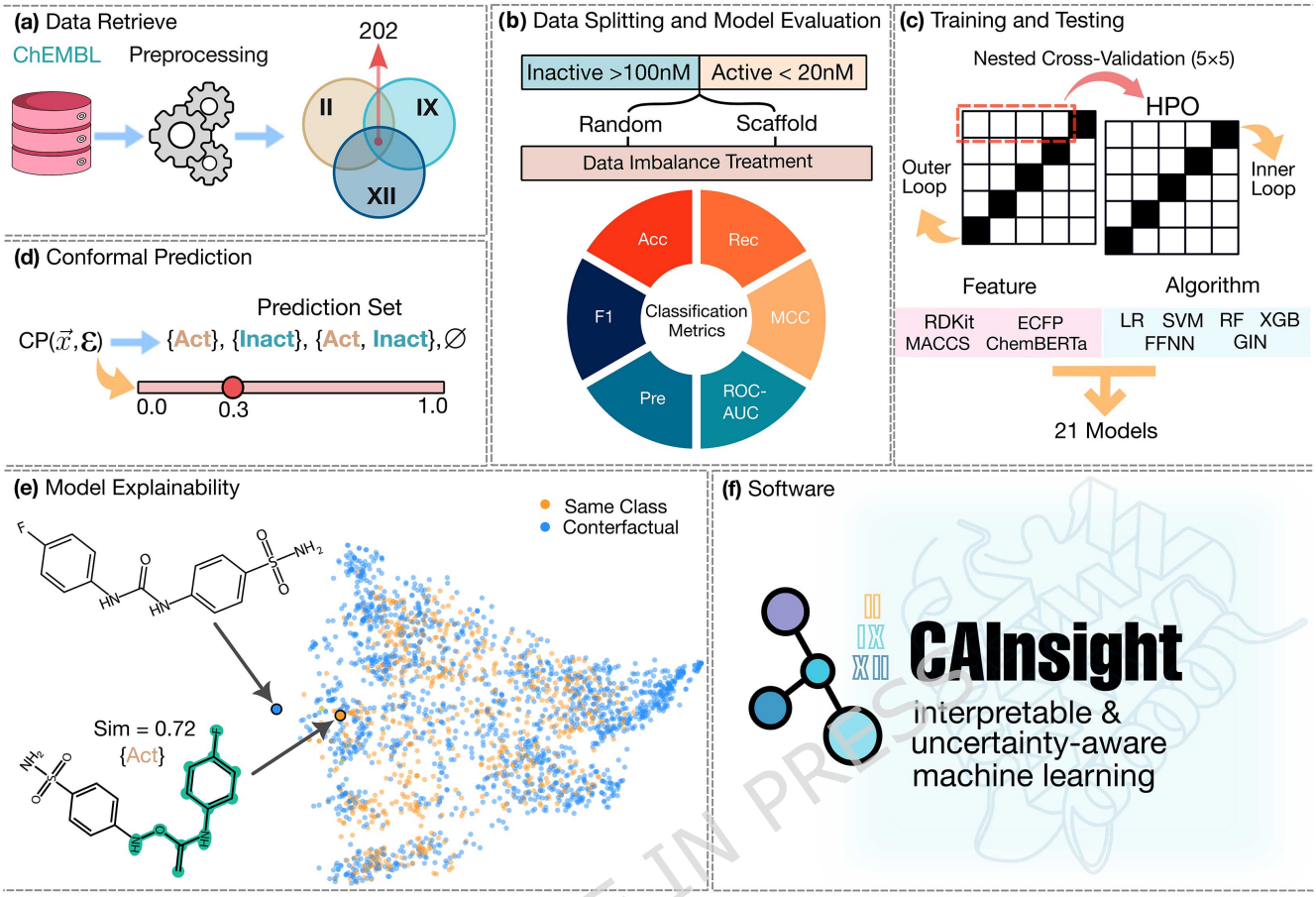

Fig. 1. Workflow of the current study. (a) Retrieving data from ChEMBL and preprocessing. (b) Random and scaffold-based data split and evaluation metrics. (c) Nested cross-validation for hyperparameter optimization, training, and testing, alongside feature generation methods and employed algorithms. (d) Conformal prediction for uncertainty quantification. (e) Model explainability by the counterfactual method. (f) `CAInsight` software for streamlining hCA inhibitor predictions.

Classification Metrics

We measure model performance by accuracy, recall, precision, F1-score, and Matthews Correlation Coefficient (MCC)[37] (Fig. 1b), implemented in `scikit-learn` (v1.5.1)[41], and define them as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

TN, TP, FN, and FP denote true negative, true positive, false negative, and false positive, respectively. We assess model performance using the Receiver Operating Characteristic Area Under the Curve (ROC AUC), a metric that measures how well

the model distinguishes between positive and negative classes by plotting the true positive rate against the false positive rate across classification thresholds. To address dataset imbalance, arising from unequal numbers of active and inactive molecules, we apply sample weights via the "sample_weights" parameter to prevent bias toward the majority class.

Feature Generation

Effective models need informative molecular representations: fingerprint, descriptor, graph, and chemical language-based[42,43] (Fig. 1c). As no representation is superior from the outset, we test four feature generation methods[37]: `RDKit` descriptors[44], Molecular Access System (MACCS)[45], Extended Connectivity Fingerprints (2048-bit, bond diameter two)[26], and ChemBERTa embeddings[46]. We standardize only `RDKit` descriptors using "StandardScaler" from `scikit-learn`. In contrast to the `RDKit` descriptors, ChemBERTa embeddings are used without further standardization. This approach is adopted to preserve the integrity of the learned latent space, as external scaling can distort the relative distances between molecular vectors that the ChemBERTa architecture worked to establish.

Machine Learning Algorithms

To predict hCA activity (Fig. 1c), machine-learning classifier algorithms, including logistic regression (LR), support vector machine (SVM), random forest (RF), and XGBoost (XGB)[47], were used alongside deep-learning classifiers such as Feed-Forward Neural Network (FFNN)[48] and Graph Isomorphic Network (GIN)[49]. Each model was trained five times with different random seeds to reduce randomness in performance. Details of algorithms and their implementations are provided in the Algorithm section of the Supplementary Information (SI), while hyperparameters and fine-tuning procedures performed with the `Optuna` package appear in Tables S1–S6. It is worth mentioning that we opted for independent binary classification models for each isoform due to the significant sparsity of bioactivity data across hCA II, IX, and XII in existing databases. Since most compounds are evaluated against only one isoform, separate models allowed us to utilize the maximum available experimental data for each target.

Conformal Prediction

Informed decisions—whether to accept or reject a prediction—require an understanding of the prediction's reliability. Conformal prediction (CP) addresses this need by producing prediction regions (PRs) instead of single-point ($\hat{y}$) prediction in both regression and classification settings (see Fig. 1d)[50]. Each PR is a set that contains the true label (y) of a test instance with probability at least $1-\varepsilon$. In binary classification, where $y \in \{\text{Active}, \text{Inactive}\}$, a PR may be $\{\text{Active}\}$, $\{\text{Inactive}\}$, $\{\text{Active}, \text{Inactive}\}$, or the empty set $\emptyset$. A predictor is valid if the proportion of true labels falling outside the prediction region does not exceed ε. In binary classification, efficiency is measured by the number of prediction sets that include both labels, where a lower number indicates better efficiency[51]. Detailed implementation procedures are provided in the Conformal Predictor section of the SI.

Counterfactual Explainability

Counterfactual explanations use a model-agnostic approach to identify minimal structural changes that alter a prediction[52] (Fig. 1e). Unlike feature attribution methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), which quantify the contribution of each feature[53], counterfactuals provide a more intuitive and actionable insight by revealing the specific modifications needed to reverse a model's decision[52]. In the context of activity prediction, they highlight how structural changes affect molecular activity. Because counterfactual is a model-agnostic approach, it treats the underlying algorithm (e.g., Support Vector Machine) as a black box and focuses solely on the relationship between input perturbations and output changes. This allows for a robust interpretation that is not tied to the specific complexity of the model. It is necessary to mention that counterfactual explanations are model-dependent and may be sensitive to the chosen activity thresholds; they should therefore be interpreted as local rationalizations of model behavior rather than as universal chemical mechanisms. For more information, refer to the Counterfactual Explainability section of the SI.

Results and discussion

We first build a curated dataset of three hCA isoforms using records from ChEMBL and preprocess the entries to ensure data quality (see Database Retrieval subsection). After filtering, 4479, 2776, and 2057 molecules are preserved for hCA II, hCA IX, and hCA XII isoforms, respectively. Table S7 lists K_i, K_d, and IC50 values for each isoform. The kernel density estimate in Fig. 2 shows similar activity distributions across isoforms, though hCA II and hCA IX span a broader range. Table S8 provides the detailed binding affinity statistics.

We compute molecular weight, hydrogen bond donors and acceptors, QED, LogP, and rotatable bonds using the `Datamol` package for each molecule across the three isoforms (Fig. S1, Tables S9–S11). The calculated properties remain consistent across isoforms. To evaluate druglikeness, we apply Lipinski's rule of five (Ro5)[54], which assesses whether a molecule can be absorbed and distributed in the human body. According to Ro5, a compound can be orally active if it has no more than five

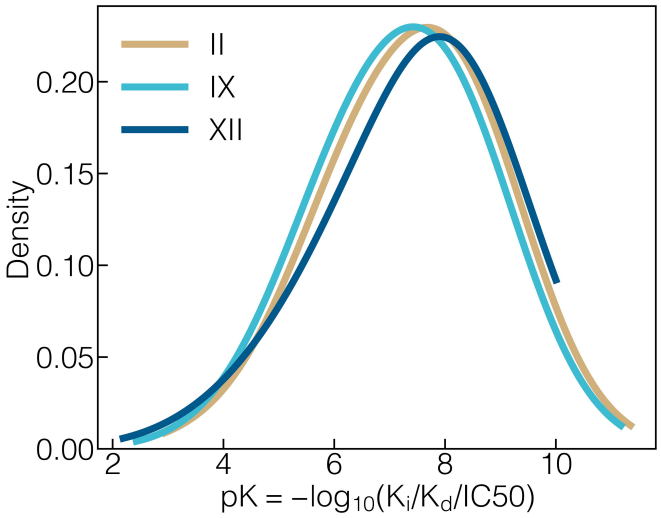

Fig. 2. Kernel density estimation (KDE) of activity values for II, IX, and XII isoforms.

hydrogen bond donors, ten hydrogen bond acceptors, a molecular mass under 500 Da, and a LogP of five or less. Approximately 66.9%, 72.4%, and 72.1% of molecules associated with isoforms II, IX, and XII satisfy the Ro5 criteria. These results show that our curated database is relevant to drug discovery.

We examine how molecules from isoforms II, IX, and XII are distributed within chemical space by employing the `MolCompass` package (v1.1.2)[55], which applies the t-distributed Stochastic Neighbor Embedding (t-SNE) reduction algorithm[56] trained on ChEMBL molecular structures. Fig. S2 shows that molecules from the three isoforms cluster in similar regions of chemical space. This observation, supported by molecular property analyses, underscores the challenge of distinguishing among ligands for these isoforms.

The aim of this study is to develop three distinct binary classification algorithms to detect active or inactive ligands for each hCA isoform. Therefore, selectivity is not modeled directly, but is instead derived by comparing the outputs of the three independent models. Reliable prediction requires careful attention to four methodological steps: dataset splitting (training, validation, and testing), handling data imbalance, optimizing hyperparameters, and implementing a thorough model evaluation procedure.

We split the dataset using both random and scaffold-based sampling (see Dataset Split section). Addressing data imbalance is equally essential, as classification algorithms trained on imbalanced datasets are prone to being biased toward the majority class (e.g., in binary setting, the class with the greater number of data instances). To reduce this bias, we apply random oversampling of the minority class using the `imbalanced-learn` package (v0.13.0)[57].

The hyperparameters are optimized by the `Optuna` framework with its default settings. For each algorithm, we define the hyperparameters as the objective function and set the number of trials to 20. Tables S2–S6 list the hyperparameter ranges.

Model performance is evaluated using 5 × 5-fold nested cross-validationas, as suggested by Ash et al.[58], to ensure rigor and reduce bias. The procedure uses an inner loop for hyperparameter tuning and an outer loop to estimate performance on unseen data. In the outer loop, the dataset is split into five mutually exclusive folds. Each iteration reserves one fold for testing while the remaining four form the training set. The training set enters the inner loop, where 5-fold cross-validation tunes hyperparameters. Models train on each inner training subset and are evaluated on the corresponding validation subset. The configuration that minimizes the average validation error across inner folds is selected. With optimal hyperparameters, the model is retrained on the outer training set and evaluated on the test fold. This process repeats five times, and the average performance is reported to reduce randomness. Both random and scaffold-based sampling strategies are applied throughout.

We build 20 models by combining four feature representations—molecular descriptors, MACCS keys, ECFP, and ChemBERTa—and five ML and DL algorithms (i.e., RF, LR, SVM, XGB, and FFNN). Further, we use GIN as a well-established algorithm from the GNN family. Finally, the model's performance is assessed using Precision, Recall, Accuracy, F1 score, ROC-AUC, and MCC metrics (see Classification Metrics subsection for more details).

Tables S12–S17 report the mean and standard deviation of evaluation metrics for every algorithm–feature pair across three isoforms and two split strategies: random and scaffold-based. As seen in Fig. 3, random splits, compared with scafold splits, produce higher scores but overestimate performance and thus reduce the reliability, as training and test sets share similar structures. This is because, the scaffold splits avoid the overlap, by ensuring chemical dissimilarity between training and test

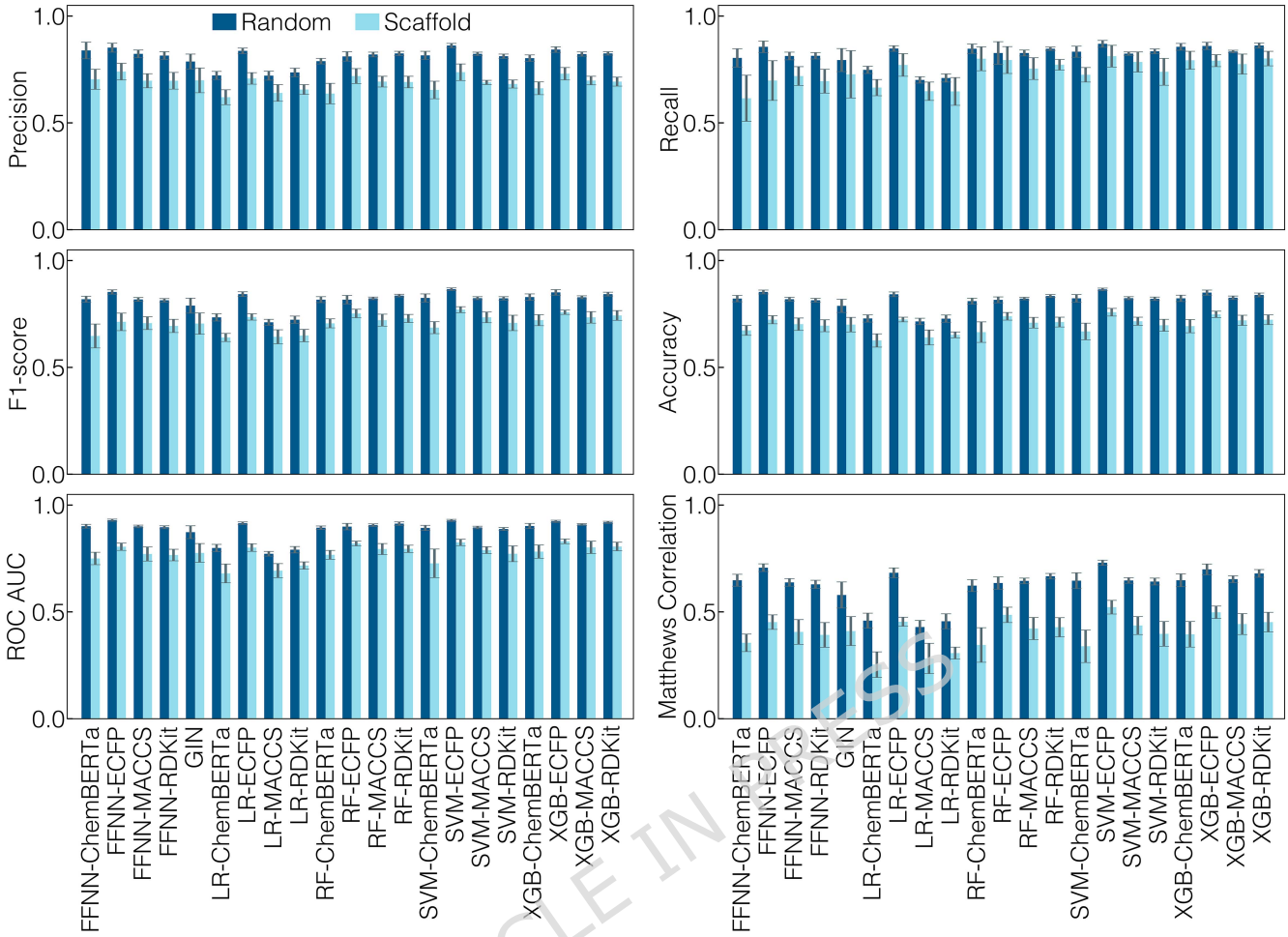

Fig. 3. Comparison of Random and Scaffold splits performance across different metrics and algorithms/features for hCA II.

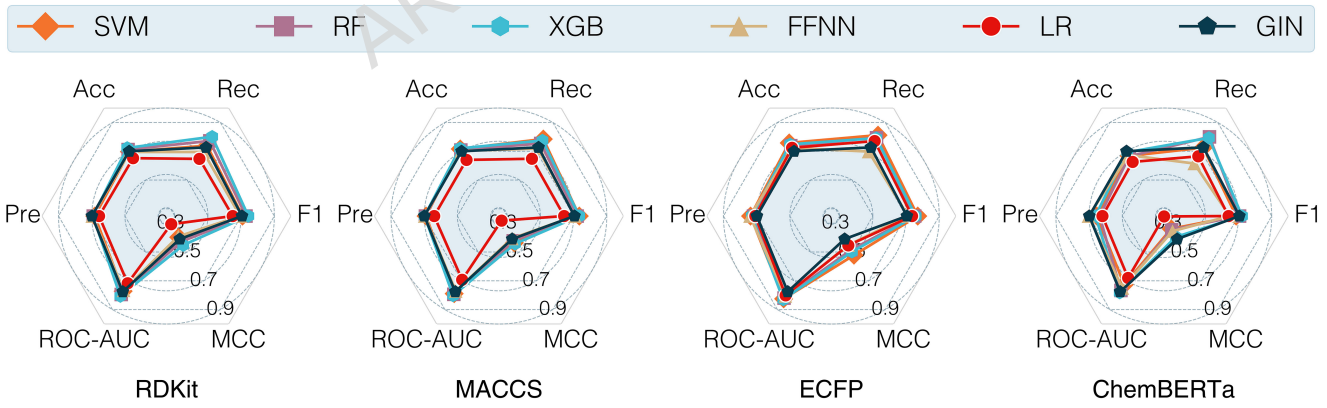

Fig. 4. The performance spider plots for the hCA II. ACC, Rec, MCC, ROC-AUC, and Pre are Accuracy, Recall, Matthews Correlation Coefficient, Receiver Operating Characteristic Area Under the Curve, and Precision, respectively.

sets, offering a more realistic measure of generalizability. We therefore use scaffold splits to select the best model. Figs S3 and S4 compare random and scaffold splits for isoforms IX and XII across all metrics, algorithms, and feature sets.

Under the scaffold-based evaluation, the SVM-ECFP model ranks first in most metrics for all three isoforms. For isoform II, SVM-ECFP achieves the highest scores in Recall (0.812 ± 0.051), Accuracy (0.758 ± 0.018), F1-score (0.770 ± 0.014), and MCC (0.522 ± 0.032). It is slightly outperformed in Precision by FFNN-ECFP (0.740 ± 0.039) and in ROC-AUC by

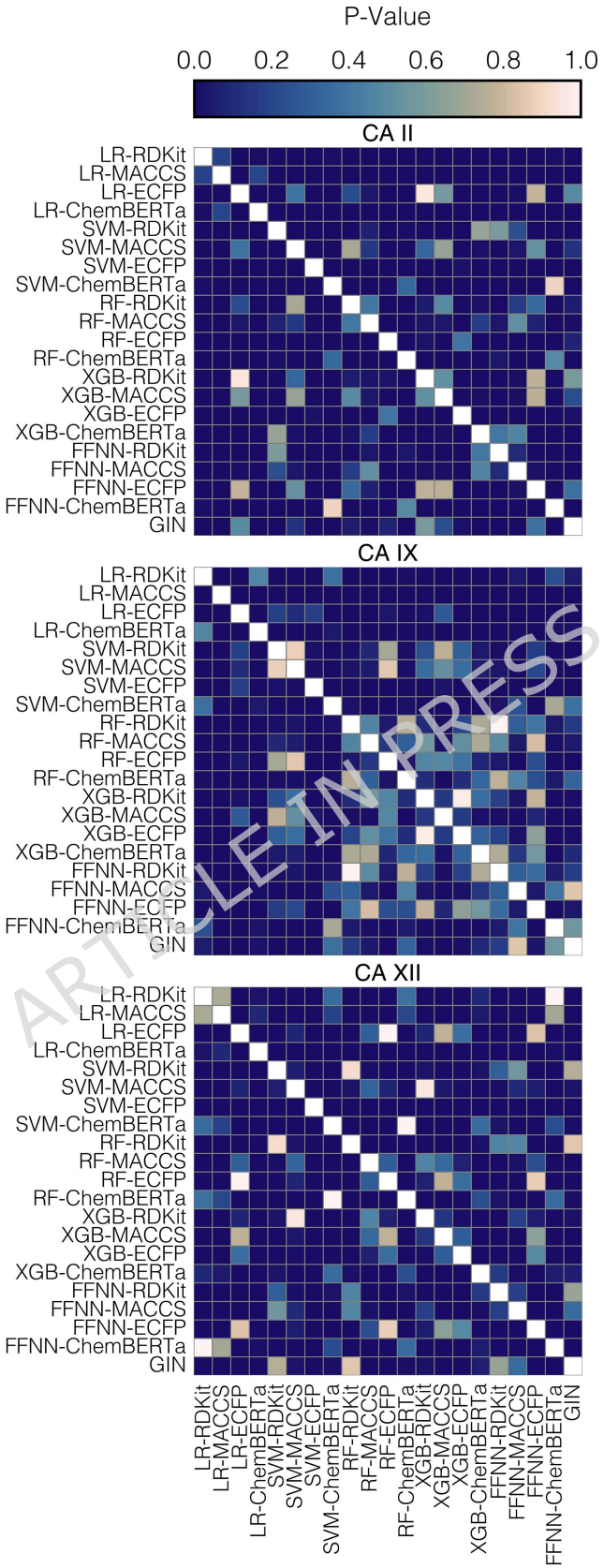

Fig. 5. Heatmap illustrating p-values from McNemar's test, highlighting statistically significant differences in performance between algorithms-features for three isoforms.

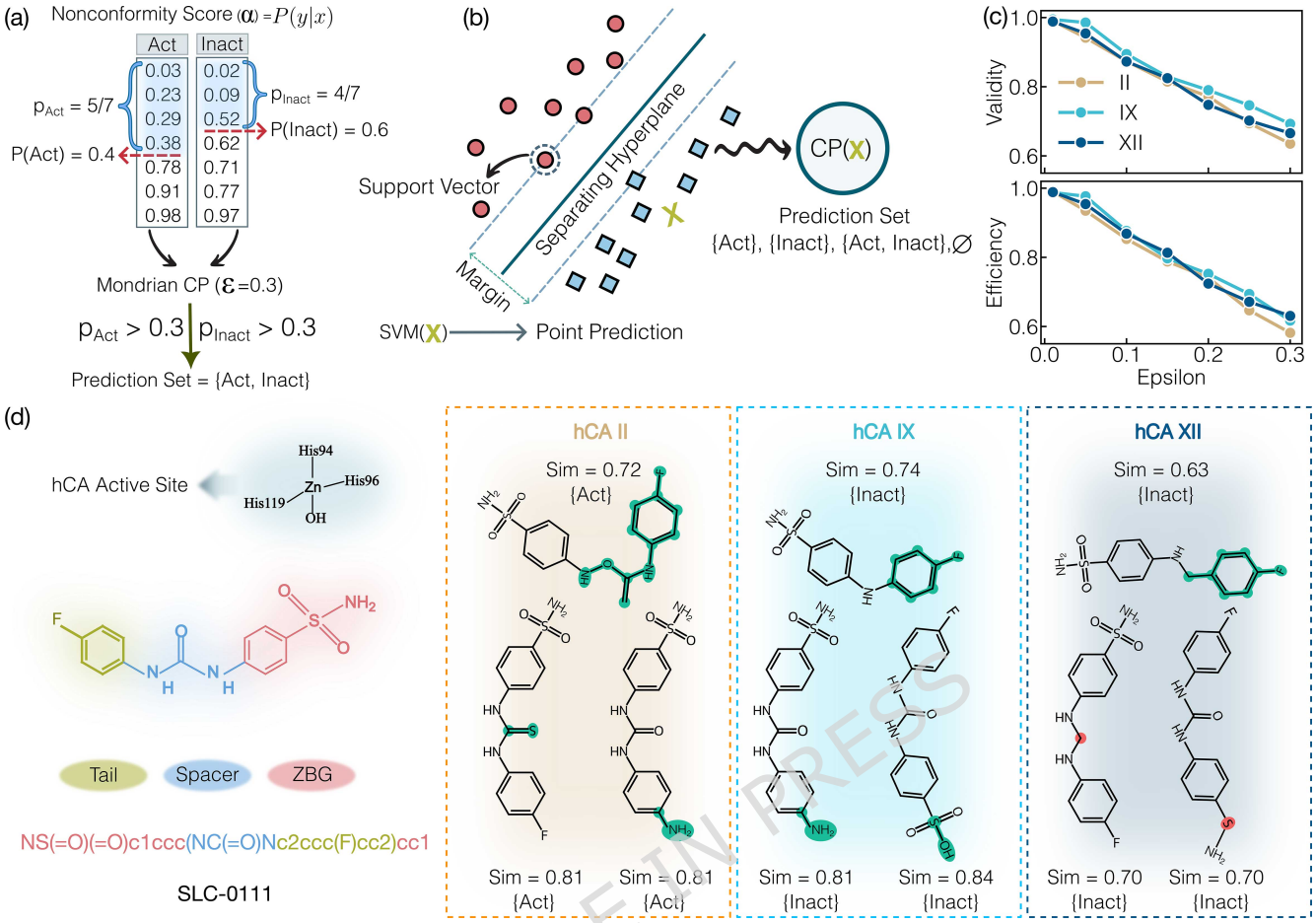

Fig. 6. Schematic of the Mondrian Conformal Prediction process. (a) p_{Act} and p_{Inact} are computed for active and inactive classes, respectively, based on the new sample's nonconformity score. The class is considered if its p variable exceeds ε. (b) An SVM classifier is used to generate the nonconformity scores required for the Mondrian CP approach. The model can produce prediction sets that are single ("active" or "inactive"), multiple ("active" and "inactive"), or empty (∅). (c) Validity and efficiency plots based on ε for hCA II, hCA IX, and hCA XII. (d) Counterfactual explainability reveals influential parts of the SLC-0111 molecule as inhibitors for hCA IX and hCA XII.

XGB-ECFP (0.830 ± 0.011). For isoform IX, SVM-ECFP also yields the highest values for Recall (0.752 ± 0.101), Accuracy (0.720 ± 0.040), F1-score (0.726 ± 0.054), ROC-AUC (0.795 ± 0.039), and MCC (0.446 ± 0.078), while only Precision is surpassed by LR-ECFP (0.713 ± 0.064). A similar trend is observed for isoform XII, where SVM-ECFP again outperforms other models in most metrics, including Recall (0.896 ± 0.014) and F1-score (0.809 ± 0.030), while FFNN-ECFP achieves the highest Precision (0.750 ± 0.037). These results suggest that SVM is the most effective algorithm and ECFP is the most informative feature representation on average. Spider plots illustrating the performance of the classification algorithms across different feature types for hCA II are shown in Fig. 4, and for isoforms IX and XII in Figs S5 and S6, respectively.

The consistent outperformance of SVM-ECFP over GIN and ChemBERTa's embedding is a finding that aligns with recent studies in the field when the dataset volume is small to medium[59]. We employed GIN because popular models, such as Graph Convolutional Networks (GCN), can not distinguish between non-isomorphic graph structures[60]. In contrast, GIN was specifically designed to be as powerful as the Weisfeiler-Lehman (WL) graph isomorphism test[49]. However, GIN, similar to other Neural Networks (NNs), is inherently data-hungry. In our study, the sample size may not reach the adequate threshold for deep learning models to decode complex latent relationships compared to fixed ECFP fingerprints.

We also opted for ChemBERTa embedding because of its excellence and widespread usage in this discipline[46]. The superior performance of ECFP over ChemBERTa embeddings is related to the nature of these representations. ECFP fingerprints represent a pre-computed feature vector, where structural information like the presence of rings, bonds, and specific functional groups is hard-coded. This explicit representation allows classical machine learning models to identify key features immediately. In contrast, while ChemBERTa is pre-trained on millions of molecules to capture general chemical syntax, its output 384-

dimensional embeddings require the downstream classifier to decode complex latent relationships to predict a bioactivity. With a small-to-medium dataset, the models lack sufficient data to learn the intricate mapping between latent representation and bioactivity[46].

Besides performance-based comparison, we apply McNemar's test as a statistical method designed to evaluate differences in paired binary outcomes[61]. McNemar's test determines whether the proportion of successes and failures differs significantly between two classifiers. During assessment, we set the significance threshold to $p < 0.01$. The results, presented as a heatmap in Fig. 5 and depicted in Tables S18-S21, support the findings from the performance metrics. Overall, both the performance metrics and the McNemar test consistently identify SVM-ECFP as the best-performing model with statistically significant differences.

To evaluate the practical significance of the SVM-ECFP model, we compared its performance with the next best models across all metrics. For isoform II (Table S13), SVM-ECFP attains the highest F1-score, 0.770 ± 0.014, followed closely by XGB-ECFP at 0.758 ± 0.010. Although the difference in performance is modest, SVM offers inherent simplicity and greater computational efficiency than XGB. This combination of efficiency and strong predictive accuracy makes SVM-ECFP a compelling balance between performance and resource cost. Similar patterns emerge across other isoforms (Tables S15 and S17). In this context, SVM-ECFP not only demonstrates statistically significant superiority but also provides a practical and parsimonious choice for ligand classification.

Table 1. Mean and standard deviation (in parenthesis) values for validity and efficiency from conformal predictions for three isoforms.

	hCA II		hCA IX		hCA XII	
Epsilon	Validity	Efficiency	Validity	Efficiency	Validity	Efficiency
0.01	0.990(0.001)	0.990(0.001)	0.995(0.000)	0.988(0.001)	0.989(0.005)	0.989(0.005)
0.05	0.942(0.002)	0.935(0.002)	0.986(0.000)	0.977(0.000)	0.954(0.003)	0.954(0.003)
0.10	0.874(0.003)	0.853(0.004)	0.894(0.002)	0.876(0.002)	0.873(0.010)	0.868(0.009)
0.15	0.815(0.004)	0.788(0.004)	0.829(0.001)	0.797(0.001)	0.825(0.013)	0.813(0.014)
0.20	0.772(0.003)	0.741(0.004)	0.790(0.002)	0.752(0.002)	0.748(0.011)	0.724(0.012)
0.25	0.695(0.006)	0.647(0.006)	0.746(0.001)	0.694(0.001)	0.702(0.012)	0.671(0.014)
0.30	0.635(0.003)	0.582(0.003)	0.693(0.001)	0.617(0.001)	0.666(0.016)	0.631(0.018)

So far, our results indicate that SVM-ECFP achieves the best performance across all three isoforms. Even so, every model has a limited domain of applicability due to the finite scope of the training data. Instances that differ substantially from the training set tend to produce higher prediction errors and greater uncertainty. To address this limitation, we apply the Mondrian Conformal Prediction (CP) framework to SVM-ECFP for all isoforms, enabling rigorous uncertainty quantification and more informed decision-making. The general workflow of Mondrian CP is shown in Fig. 6a and b (see the Conformal Predictor section of the SI for further details).

As noted, the key characteristics of a conformal predictor are validity and efficiency. Table 1 and Fig. 6c present detailed results for these criteria across different significance levels (ε) for the three hCA isoforms. Validity closely matches the expected coverage rate of $1 - \varepsilon$, confirming the theoretical guarantees of the conformal prediction framework. For instance, at $\varepsilon = 0.01$, the models achieved validity scores of 0.990, 0.995, and 0.989, respectively. We also observe the anticipated trade-off between ε and efficiency: as ε increases, prediction sets become larger and less precise. In hCA II, for example, increasing ε from 0.01 to 0.3 reduces efficiency from 0.990 to 0.582.

Although the preceding results confirm the reliability of our ML models, their black-box nature requires explainability methods to clarify their reasoning. To this end, we augment our models with a model-agnostic counterfactual explainability approach (see Counterfactual Explainability subsection). This method probes the chemical space surrounding a given molecule by generating structurally similar analogs with minimal modifications, thereby revealing the basis of the model's predictions. To illustrate the utility of this approach, we examine the clinical-stage sulfonamide drug candidate SLC-0111[62] (ChEMBL ID CHEMBL1615281; Fig. 6d). SLC-0111 inhibits hCA IX and hCA XII but not isoform II[63,64], a selectivity that is crucial for its anticancer efficacy in hypoxic environments.

SLC-0111 inhibits its target by coordinating its deprotonated sulfonamide group with the catalytic zinc ion, thereby displacing the zinc-bound hydroxide. This core interaction is shared across hCA II, IX, and XII. Selectivity, however, stems from the compound's tail group, which forms specific hydrogen bonds and van der Waals interactions with unique residues in the broader active sites of hCA IX and XII, resulting in high-affinity, stable binding[63,64].

For each isoform, we generate three counterfactual molecules using the `examol` package (v3.3.0)[52]. The results (Fig. 6d)

show that modifications in the tail region strongly affect molecular selectivity, underscoring the tail group as a key structural determinant for differentiating inhibitor activity among hCA isoforms II, IX, and XII. Our findings reveal that our models align with well-established mechanistic knowledge in medicinal chemistry and could independently recapture these known structural determinants. The corresponding chemical space for selectivity prediction is presented in Figs S7–S9 for the three isoforms.

To validate the model's ability to distinguish subtle differences among chemical scaffolds, we analyzed non-selective compounds. To this end, we chose CHEMBL1338403 and CHEMBL120886, which are active and inactive against all hCA isoforms, respectively. The analysis of CHEMBL1338403 highlights the role of the sulfonamide group and hydrophobic tail in enhancing potency, while CHEMBL120886 shows that its benzoic acid scaffold's chlorine atoms and carboxylate group hinder activation (see Figs S10-S15 for counterfactual chemical spaces).

While these examples demonstrate the model's discriminative power, it's important to note that the counterfactual method is appropriate for local explanations, and applying it to an entire dataset to identify a global pattern remains an ongoing challenge in explainable AI. Although these counterfactual patterns align with known medicinal chemistry principles, they reflect the decision boundaries of our specific trained model and dataset. Interpretations could shift under different labeling schemes, underscoring that such explanations are model-aware rather than purely mechanistic.

Although sharing code has become standard practice (see Data Availability), such resources are often inaccessible to users with limited programming experience. To address this, we introduce `CAInsight`, a user-friendly Graphical User Interface (GUI) for exploring the tools and models developed in this study (Figs 1f and S16). `CAInsight` is implemented with `streamlit` (v1.25.0) and packaged via `Mamba`, a cross-platform package manager that simplifies installation (see the GitHub repository for details). The interface integrates three core tools: SVM-ECFP classifiers for all three isoforms, conformal predictors for uncertainty quantification and informed decision-making, and counterfactual explainability for interpreting model predictions. The workflow is streamlined, requiring only the molecule's SMILES and a single click on the Run button to generate results.

Conclusion

This work addresses key limitations in the computational prediction of hCA inhibitors. We show that an SVM model using ECFP provides an optimal approach for predicting the binding affinity of inhibitors for cancer-related isoforms IX and XII, as well as the off-target isoform II. Notably, improved performance arises primarily from careful data preprocessing and the handling of class imbalance, rather than from algorithmic complexity. Our framework offers a powerful and interpretable tool for identifying new inhibitors by combining conformal prediction for reliability with counterfactuals for interpretability. It is important to note that our proposed classifiers predict only the activity or binding of a compound to specific isoforms, and that selectivity is inferred subsequently by comparing predictions across the three independent classifiers. While this work demonstrates a proof-of-concept through specific case studies like SLC-0111, we recognize some limitations and the need for a more comprehensive selectivity analysis across broader chemical spaces. Additionally, as a methodological limitation, we note that the chosen activity thresholds (active < 20 nM, inactive > 100 nM) were not subjected to a formal sensitivity analysis. Future work should explore how threshold selection affects model performance and interpretation. Furthermore, the aggregation of K_i, K_d, and IC_{50} values—while enabling a more chemically diverse training set—introduces assay-specific noise that our binary classification framework partially mitigates but does not fully eliminate. Future studies with larger, homogeneous affinity datasets would help refine the potency predictions and robustness of our models. All tools and models are made accessible through the `CAInsight` software, enabling broader exploration. We anticipate that these models and tools will prove valuable for virtual screening of large chemical spaces[65] and as scoring functions in generative deep learning approaches[66] to identify potential hit molecules. By facilitating more effective in silico filtering prior to synthesis and experimental validation, these resources can accelerate the discovery of safer therapeutic candidates.

Funding

This work is based upon research funded by Iran National Science Foundation (INSF) under project No.4037187.

References

1. Manzari, M. T. *et al.* Targeted drug delivery strategies for precision medicines. *Nat. Rev. Mater.* **6**, 351–370 (2021).
2. Srinivasarao, M. & Low, P. S. Ligand-Targeted Drug Delivery. *Chem. Rev.* **117**, 12133–12164 (2017).
3. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
4. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).

5. Pan, Y., Liu, L., Mou, X. & Cai, Y. Nanomedicine Strategies in Conquering and Utilizing the Cancer Hypoxia Environment. *ACS Nano* **17**, 20875–20924 (2023).

6. Liu, J.-n., Bu, W. & Shi, J. Chemical Design and Synthesis of Functionalized Probes for Imaging and Treating Tumor Hypoxia. *Chem. Rev.* **117**, 6160–6224 (2017).

7. D'Ambrosio, K. *et al.* Multiple Binding Modes of Inhibitors to Human Carbonic Anhydrases: An Update on the Design of Isoform-Specific Modulators of Activity. *Chem. Rev.* **125**, 150–222 (2025).

8. Baroni, C. *et al.* Lasamide, a Potent Human Carbonic Anhydrase Inhibitor from the Market: Inhibition Profiling and Crystallographic Studies. *ACS Med. Chem. Lett.* **15**, 1749–1755 (2024).

9. Eldehna, W. M. *et al.* Benzofuran-Based Carboxylic Acids as Carbonic Anhydrase Inhibitors and Antiproliferative Agents against Breast Cancer. *ACS Med. Chem. Lett.* **11**, 1022–1027 (2020).

10. Peerzada, M. N. *et al.* Discovery of Novel Hydroxyimine-Tethered Benzenesulfonamides as Potential Human Carbonic Anhydrase IX/XII Inhibitors. *ACS Med. Chem. Lett.* **14**, 810–819 (2023).

11. Kciuk, M. *et al.* Targeting carbonic anhydrase IX and XII isoforms with small molecule inhibitors and monoclonal antibodies. *J. Enzyme Inhib. Med. Chem.* **37**, 1278–1298 (2022).

12. Mishra, C. B., Tiwari, M. & Supuran, C. T. Progress in the development of human carbonic anhydrase inhibitors and their pharmacological applications: Where are we today? *Med. Res. Rev.* **40**, 2485–2565 (2020).

13. Nada, H., Meanwell, N. A. & Gabr, M. T. Virtual screening: hope, hype, and the fine line in between. *Expert Opin. Drug Discovery* **20**, 145—-162 (2025).

14. Weissenow, K. & Rost, B. Are protein language models the new universal key? *Curr. Opin. Struct. Biol.* **91**, 102997 (2025).

15. Rayka, M., Mirzaei, M., Farnoosh, G. & Latifi, A. M. Investigating Enzyme Biochemistry by Deep Learning: A Computational Tool for a New Era. *J. Comput. Biophys. Chem.* **23**, 781–799 (2024).

16. Hann, M. M. & Keserű, G. M. The continuing importance of chemical intuition for the medicinal chemist in the era of Artificial Intelligence. *Expert Opin. Drug Discovery* **20**, 137–140 (2025).

17. Nobel Prize in Chemistry 2024 (2024). URL https://www.nobelprize.org/prizes/chemistry/2024/press-release.

18. Pitt, W. R. *et al.* Real-World Applications and Experiences of AI/ML Deployment for Drug Discovery. *J. Med. Chem.* **68**, 851–859 (2025).

19. Wang, H. Prediction of protein–ligand binding affinity via deep learning models. *Briefings Bioinf.* **25**, bbae081 (2024).

20. Schapin, N., Majewski, M., Varela-Rial, A., Arroniz, C. & Fabritiis, G. D. Machine learning small molecule properties in drug discovery. *Artificial Intelligence Chemistry* **1**, 100020 (2023).

21. Zhou, G. *et al.* An artificial intelligence accelerated virtual screening platform for drug discovery. *Nat. Commun.* **15**, 1–14 (2024).

22. Li, L. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat. Commun.* **14**, 1–12 (2023).

23. Sapoval, N. *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**, 1–12 (2022).

24. Galati, S. *et al.* Predicting Isoform-Selective Carbonic Anhydrase Inhibitors via Machine Learning and Rationalizing Structural Features Important for Selectivity. *ACS Omega* **6**, 4080–4089 (2021).

25. Kim, S. *et al.* PubChem 2019 update. *Nucleic Acids Res.* **51**, D1102–D1109 (2019).

26. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

27. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).

28. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (" O'Reilly Media, Inc.", 2022).

29. Tinivella, A., Pinzi, L. & Rastelli, G. Prediction of activity and selectivity profiles of human Carbonic Anhydrase inhibitors using machine learning classification models. *J. Cheminf.* **13**, 1–15 (2021).

30. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).

31. Li, J. *et al.* Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction (2024). URL https://arxiv.org/html/2308.09639v2. [Online; accessed 3. Jul. 2024].

32. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).

33. Tang, J. *et al.* Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis. *J. Chem. Inf. Model.* **54**, 735–743 (2014).

34. Contributors, D. Datamol: Molecular processing made easy (2025). URL https://github.com/datamol-io/datamol.

35. Rdkit: Open-source cheminformatics. https://www.rdkit.org (2023). Version 2023.9.5.

36. Identification of Novel Carbonic Anhydrase IX Inhibitors Using High-Throughput Screening of Pooled Compound Libraries by DNA-Linked Inhibitor Antibody Assay (DIANA) (2020). [Online; accessed 28. Dec. 2025].

37. Deng, J. *et al.* A systematic study of key elements underlying molecular property prediction. *Nat. Commun.* **14**, 1–20 (2023).

38. Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).

39. Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).

40. Ramsundar, B. *et al. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more* (O'Reilly Media, Inc., 2019).

41. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

42. Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.* **12**, e1603 (2022).

43. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminf.* **12**, 1–22 (2020).

44. Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **37**, 1376–1382 (2021).

45. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).

46. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* (2020). 2010.09885.

47. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *ACM Conferences*, 785–794 (Association for Computing Machinery, New York, NY, USA, 2016).

48. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*, vol. 1 (MIT press Cambridge, 2016).

49. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? arxiv preprint arxiv: 181000826. *Published online* (2018).

50. Arvidsson McShane, S. *et al.* CPSign: conformal prediction for cheminformatics modeling. *J. Cheminf.* **16**, 75–17 (2024).

51. Norinder, U., Carlsson, L., Boyer, S. & Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **54**, 1596–1603 (2014).

52. Wellawatte, G. P., Seshadri, A. & White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.* **13**, 3697–3705 (2022).

53. Wu, Z. *et al.* From Black Boxes to Actionable Insights: A Perspective on Explainable Artificial Intelligence for Scientific Discovery. *J. Chem. Inf. Model.* **63**, 7617–7627 (2023).

54. Pollastri, M. P. Overview on the Rule of Five. *Curr. Protoc. Pharmacol.* **49**, 9.12.1–9.12.8 (2010).

55. Sosnin, S. MolCompass: multi-tool for the navigation in chemical space and visual validation of QSAR/QSPR models. *J. Cheminf.* **16**, 1–13 (2024).

56. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008). URL https://jmlr.org/papers/v9/vandermaaten08a.html.

57. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017). URL http://jmlr.org/papers/v18/16-365.

58. Ash, J. R. *et al.* Practically significant method comparison protocols for machine learning in small molecule drug discovery. *ChemRxiv* (2024).

59. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).

60. Dwivedi, V. P. *et al.* Benchmarking Graph Neural Networks. *arXiv* (2020). 2003.00982.

61. Pembury Smith, M. Q. R. & Ruxton, G. D. Effective use of the McNemar test. *Behav. Ecol. Sociobiol.* **74**, 133–9 (2020).

62. McDonald, P. C. *et al.* A Phase 1 Study of SLC-0111, a Novel Inhibitor of Carbonic Anhydrase IX, in Patients With Advanced Solid Tumors. *Am. J. Clin. Oncol.* **43**, 484 (2020).

63. Pacchiano, F. *et al.* Ureido-Substituted Benzenesulfonamides Potently Inhibit Carbonic Anhydrase IX and Show Antimetastatic Activity in a Model of Breast Cancer Metastasis. *J. Med. Chem.* **54**, 1896–1902 (2011).

64. Williams, K. J. & Gieling, R. G. Preclinical Evaluation of Ureidosulfamate Carbonic Anhydrase IX/XII Inhibitors in the Treatment of Cancers. *Int. J. Mol. Sci.* **20**, 6080 (2019).

65. Thaingtamtanha, T., Ravichandran, R. & Gentile, F. On the application of artificial intelligence in virtual screening. *Expert Opin. Drug Discovery* **20**, 845–857 (2025).

66. Gangwal, A. & Lavecchia, A. Unleashing the power of generative ai in drug discovery. *Drug Discov. Today* **29**, 103992 (2024). URL https://www.sciencedirect.com/science/article/pii/S135964462400117X.

Acknowledgements

The authors acknowledge the support and resources from the Center for High-Performance Computing (SARMAD) at Shahid Beheshti University of Iran. This work is based upon research funded by Iran National Science Foundation (INSF) under project No.4037187.

Author contributions

SSN conceptualized the study, MR and MSQ contributed equally to this study by conducting the experiments, and writing the manuscript. SSN revised the manuscript. All authors read the manuscript, provided feedback and eventually approved it in its final form.

Data availability

The data used for this project has been retrieved from the ChEMBL database (https://www.ebi.ac.uk/chembl/). All codes to reproduce the results, figures, and installing `CAInsight` are available at https://github.com/miladrayka/hca_ml.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material.