ACS Chapter Nigeria Book of Proceedings 15th Annual Symposium
ISBN: 978-978-68-0485-9 DOI: https://doi.org/10.55455/acsnigeria.1.2.149-154

Stacked Ensemble Machine Learning for Human Carbonic Anhydrase II Binding Affinity Prediction

[1]Miracle Olatilewa Olapade, [2]Olalekan Cosmas Ogundola, and [1]Dotun Solomon Akeju.
[1]Department of Chemistry, University of Ibadan, Ibadan, Nigeria
[2]Department of Chemistry, Federal University, Oye-Ekiti, Nigeria **Corresponding Author's email:** olapademiracleo@gmail.com

ABSTRACT

Many diseases, such as glaucoma, epilepsy, and cancer, are associated with carbonic anhydrase II. This makes it an important therapeutic target for these diseases. A stacked ensemble machine learning model was built to predict the binding affinity of ligands with CA II. The dataset used consists of 6,530 compounds with experimental Ki values from ChEMBL. Each molecule was represented by a set of 1,420 molecular descriptors, including Morgan fingerprints, MACCS keys, and RDKit 2D descriptors, which were refined to 1,320 features through different feature selection procedures. A stacked ensemble model which makes use of LightGBM, ExtraTrees, and a Multi-Layer Perceptron (MLP) was developed, with ridge regression as the meta-learner. The model achieved a satisfactory performance on the test set, with a root mean square error (RMSE) of 0.68 pKi units and a coefficient of determination (R^2) of 0.76. SHAP (SHapley Additive exPlanations) analysis of the best-performing model provided important interpretability. The method identified some specific molecular substructures (e.g., Morgan_833), key pharmacophoric elements (MACCS_84, MACCS_33), and functional groups (e.g., primary amines) as the most impactful drivers of binding affinity. The outcome of the study aligns with established structure-activity relationships for CA II inhibitors; this validates the model's decision-making process. This work provides more than a tool for virtual screening but also offers interpretable insights to guide the rational design of novel CA II inhibitors.

KEYWORDS: Machine Learning, Binding Affinity, SHAP analysis, Carbonic Anhydrase II, Virtual Screening

1. INTRODUCTION

Human Carbonic Anhydrase II (hCA II) is a zinc-dependent metalloenzyme essential for physiological pH regulation and a well-established therapeutic target for conditions such as glaucoma and epilepsy, which drives the development of inhibitors like sulfonamides.[1,2,3] While experimental binding affinity measurements are resource-intensive,[4] conventional computational methods like molecular docking and dynamics are constrained by force-field approximations and high computational cost.[5,6,7,8] Machine learning (ML) offers a powerful alternative by learning structure-activity relationships directly from data, often surpassing the performance of classical scoring functions.[9]

Machine learning has been increasingly applied in carbonic anhydrase research, primarily focusing on classification tasks such as predicting inhibitor activity (active/inactive)[10] and identifying multi-target inhibitors.[11] Some studies have addressed the critical challenge of isoform selectivity, developing models to distinguish inhibitors of off-target isoforms like hCA II from therapeutic targets such as hCA IX,[12] with recent work incorporating explainable AI to elucidate the structural basis of these predictions.[13] Although these classification approaches are valuable tools for virtual screening, they offer limited utility for lead optimization, which requires quantitative potency measurements. Predicting continuous binding affinity (pKi) values represents a complex but practically important task which enables the precise ranking of compounds and also provides the avenue for quantitative structureactivity relationship analysis. To address this need, the present study developed a stacked ensemble model to improve predictive accuracy and generalization for binding affinity prediction. This approach provides both high-accuracy pKi values and interpretable insights into the molecular features governing binding affinity.

2. METHODOLOGY

2.1. Data Curation and Preprocessing

An initial dataset of 10,294 potential human Carbonic Anhydrase II (CA II) inhibitors was sourced from the ChEMBL database.[14] This raw data was rigorously curated to ensure data quality, retaining only entries with precisely defined equilibrium dissociation constant (K_i) values reported

in nanomolar (nM) units. Compounds annotated with inequality modifiers (e.g., '>', '<') were excluded.

This filtration process resulted in a refined, high-confidence dataset of 6,539 compounds, each with a

defined ChEMBL identifier, SMILES string, and exact K_i value. For each unique canonical SMILES, only the entry with the lowest reported Ki value (indicating the highest potency) was retained to represent that compound, ensuring no data leakage between training and test sets. This process yielded a final curated dataset of 6,530 compounds. The Ki (nM) values were converted to pKi, the negative logarithm of the Ki in molar units, to create a more normally distributed target variable suitable for regression modeling, using the standard transformation:

$pKi = 9 - \log_{10}(Ki\_nM)$

2.2. Molecular Feature Engineering and Selection

Molecular descriptors and fingerprints were computed for each compound using the RDKit library to numerically encode their structural and physicochemical properties. This included: (i) Morgan Fingerprints (ECFP4-like), configured with a radius of 2 and a fixed length of 1024 bits to capture atomic environments and molecular substructures; (ii) 167-bit MACCS keys, which are binary fingerprints which shows the presence of specific predefined structural fragments; and (iii) 208 RDKit 2D descriptors capturing key properties. The combination of these features resulted in an

initial high-dimensional feature matrix comprising 1,413 dimensions for each molecule and whereby after feature selection we have 1320 features for model development and evaluation.

2.3. Model Development and Evaluation

The curated dataset was partitioned using stratified random split into training set (60%), validation (20%), and test sets (20%). A stacked ensemble architecture was implemented, using three base learners (LightGBM, ExtraTrees, and MLP Regressor). Hyperparameter optimization was conducted through randomized search with 3-fold cross-validation over 25 iterations. Model performance was measured using root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) metrics. To enable interpretation of the model, SHapley Additive exPlanations (SHAP) analysis was applied to elucidate feature contributions to model predictions.[15]

3. RESULTS AND DISCUSSION

3.1. Results

3.1.1 Model Performance Evaluation

The performance of all models on both validation and test sets is summarized in Table 1. The Stacked Ensemble achieved the best performance as shown in Figure 1

Table 1. Performance metrics of models on validation and test sets

Model	Validation Set			Test Set		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Extra Trees	0.7084	0.5006	0.7436	0.6895	0.4919	0.7571

LightGBM	0.7188	0.5257	0.7418	0.7052	0.5317	0.7459
MLP Neural Network	0.8144	0.6167	0.6610	0.7781	0.5910	0.6906
Stacked Ensemble	0.6920	0.5063	0.7553	0.6815	0.5033	0.7627

On the validation set, the stacked ensemble achieved the lowest RMSE (0.6920) and highest R^2 (0.7553), outperforming all individual models. Extra Trees was the strongest base learner (RMSE = 0.7084, R^2 = 0.7436), followed closely by LightGBM (RMSE = 0.7188, R^2 = 0.7418). This performance hierarchy was maintained on the independent test set as shown in Figure 1, where the stacked ensemble further improved to RMSE = 0.6815 and R^2 = 0.7627, confirming robust generalization and the advantage of the ensemble approach.

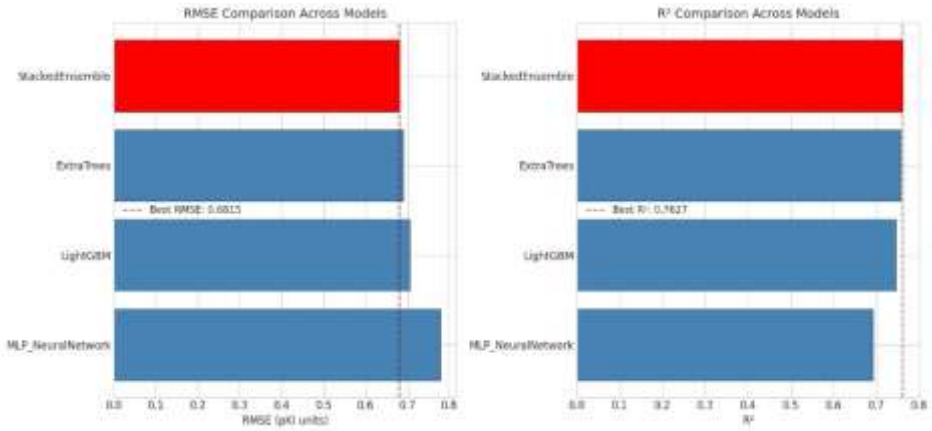

Figure 1: Comparative performance of individual models and stacked ensemble on the test set.

3.1.2 Model Interpretation via SHAP Analysis

SHAP analysis of the three individual models provided insights into binding affinity determinants.[15] The key molecular features driving predictions were identified through the SHAP analysis as shown in table 2.

Table 2. Top 5 Features by Model from SHAP Analysis

Rank	ExtraTrees	LightGBM	MLP Neural Network
1	Morgan_833 (0.3366)	SMR_VSA4 (0.4195)	fr_NH2 (0.1297)
2	MACCS_84 (0.3175)	Morgan_833 (0.1916)	SMR_VSA4 (0.0679)
3	MACCS_33 (0.0864)	MACCS_84 (0.1122)	fr_sulfonamd (0.0489)
4	fr_NH2 (0.0772)	MolLogP (0.0606)	NHOHCount (0.0258)
5	Morgan_583 (0.0501)	SPS (0.0453)	FpDensityMorgan3 (0.0227)

3.2. Discussion

This study reports the design and evaluation of a reliable predictive stacked ensemble model for predicting binding affinity (pKi) of ligands to human Carbonic Anhydrase II (CA II). The stacked

ensemble model demonstrated strong predictive performance on the test set (RMSE = 0.68, R^2 = 0.76), confirming the efficacy of this approach for QSAR modeling. Consistent with our findings, previous QSAR studies have demonstrated that stacked ensemble approaches often outperform individual models in terms of predictive accuracy.[16,17,18] The stacked ensemble approach in this study leveraged the distinct strengths of its individual models: the gradient-boosting power of LightGBM,[19] the random feature selection and the use of random thresholds to split nodes in the decision trees in ExtraTrees,[20,21] and the ability of MLPs to capture complex non-linearities and complex feature interactions[22] that may be missed by tree-based methods. The meta-learner (Ridge regression) weighted these predictions, assigning the highest weights to the tree-based models, which individually performed best.

The curation process of the molecular data was a critical factor in the model's success. The steps taken including SMILES standardization, removal of inorganic compounds, deduplication, and feature selections which are considered best practices in computational chemistry to ensure data quality and model reliability.[23] The removal of highly correlated features is crucial, as it reduces redundancy and multicollinearity, which can inflate variance and destabilize model coefficients. The model's performance on the test set (R^2 = 0.76, RMSE = 0.68) is consistent with the high standards seen in modern QSAR benchmarks. An RMSE of 0.68 log units, which corresponds to a less than 5-fold error in Ki value prediction on average, is considered highly accurate for practical applications in drug discovery, such as virtual screening and lead optimization prioritization.[24]

The SHAP analysis provided an important, experimentally actionable interpretation of the model predictions, this reveals the key structural drivers. A strong consensus across models on specific molecular features, particularly the Morgan_833 fingerprint and MACCS_84 key, suggests the identification of important substructures that are strong determinants of binding affinity. The outcome is consistent with the known structure-activity relationships of CA II inhibitors, which often rely on a zincbinding group and specific aromatic moieties that fit into the hydrophobic pocket of the enzyme. [25, 26] Hydrogen-bonding features like the primary amine count (fr_NH2) and the primary sulfonamide group (fr_sulfonamd) serves as a strong validation of the model's ability to recapitulate known medicinal chemistry, as these groups are known to coordinate the active site zinc ion. Also, the impact of properties like MolLogP and SMR_VSA4 shows the model's recognition that overall physicochemical properties are vital for optimizing ligand efficiency and bioavailability.[27] These interpretability results significantly enhance the use of the model, as they provide medicinal chemists with specific guidance on which functional groups and properties to modify in order to optimize compound affinity.

4. CONCLUSION

This study presents a stacked ensemble model and interpretation of the three individual models used as the base learners. The SHAP analysis of the three individual models shows some important features which are critical for the prediction and also provide insight for the rational design of novel CAII inhibitors.

Code and dataset are available at https://github.com/miraculinp/CAII

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

REFERENCES

(1) Tinazzi, E.; Patuzzo, G.; Lunardi, C. Autoantigens and Autoantibodies in the Pathogenesis of Sjögren's Syndrome. In *Sjogren's Syndrome*; Gerli, R., Bartoloni, E., Alunno, A., Eds.; Elsevier, 2016; pp 141– 156.

(2) O'Herin, C. B.; Moriuchi, Y. W.; Bemis, T. A.; Kohlbrand, A. J.; Burkart, M. D.; Cohen, S. M. Development of Human Carbonic Anhydrase II Heterobifunctional Degraders. *J. Med. Chem.* **2023**, *66* (4), 2789–2803. https://doi.org/10.1021/acs.jmedchem.2c01843.

(3) Supuran, C. T.; Capasso, C.; De Simone, G. Carbonic Anhydrase II as Target for Drug Design. In *Carbonic Anhydrases as Biocatalysts*; Elsevier, 2015; pp 51–90.

(4) Yang, Y. X.; Zhu, B. T. Further Exploration of the Quantitative Distance-Energy and Contact NumberEnergy Relationships for Predicting the Binding Affinity of Protein-Ligand Complexes. *Biophys. J.* **2025**, *124* (7), 1166–1177. https://doi.org/10.1016/j.bpj.2025.02.021.

(5) Mushebenge, A. G.-A.; Ugbaja, S. C.; Mbatha, N. A.; Khan, R. B.; Kumalo, H. M. Assessing the Potential Contribution of In Silico Studies in Discovering Drug Candidates That Interact with Various SARSCoV-2 Receptors. *Int. J. Mol. Sci.* **2023**, *24* (21), 15518. https://doi.org/10.3390/ijms242115518.

(6) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *9*, 1089. https://doi.org/10.3389/fphar.2018.01089.

(7) Salmaso, V.; Moro, S. Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview. *Front. Pharmacol.* **2018**, *9*, 923. https://doi.org/10.3389/fphar.2018.00923.

(8) Challapa-Mamani, M. R.; Tomás-Alvarado, E.; Espinoza-Baigorria, A.; León-Figueroa, D. A.; Sah, R.;
Rodriguez-Morales, A. J.; Barboza, J. J. Molecular Docking and Molecular Dynamics Simulations in Related to *Leishmania Donovani*: An Update and Literature Review. *Trop. Med. Infect. Dis.* **2023**, *8* (10), 457. https://doi.org/10.3390/tropicalmed8100457.

(9) Meli, R.; Morris, G. M.; Biggin, P. C. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-Based Deep Learning: A Review. *Front. Bioinform.* **2022**, *2*, 885983. https://doi.org/10.3389/fbinf.2022.885983.

(10) Tinivella, A.; Pinzi, L.; Rastelli, G. Prediction of Activity and Selectivity Profiles of Human Carbonic Anhydrase Inhibitors Using Machine Learning Classification Models. *J. Cheminform.* **2021**, *13* (1), 18. https://doi.org/10.1186/s13321-021-00499-y.

(11) Kim, M.-J.; Pandit, S.; Jee, J.-G. Discovery of Kinase and Carbonic Anhydrase Dual Inhibitors by Machine Learning Classification and Experiments. *Pharmaceuticals* **2022**, *15* (2), 236. https://doi.org/10.3390/ph15020236.

(12) Galati, S.; Yonchev, D.; Rodríguez-Pérez, R.; Vogt, M.; Tuccinardi, T.; Bajorath, J. Predicting IsoformSelective Carbonic Anhydrase Inhibitors via Machine Learning and Rationalizing Structural Features
Important for Selectivity. *ACS Omega* **2021**, *6* (5), 4080–4089. https://doi.org/10.1021/acsomega.0c06153.

(13) Kırboğa, K. K.; Işık, M. Explainable Artificial Intelligence in the Design of Selective Carbonic Anhydrase I-II Inhibitors via Molecular Fingerprinting. *J. Comput. Chem.* **2024**, *45* (18), 1530–1539. https://doi.org/10.1002/jcc.27335.

(14) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; et al. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2023**, *52* (D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004.

(15) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

(16) Grenet, I.; Merlo, K.; Comet, J.-P.; Tertiaux, R.; Rouquié, D.; Dayan, F. Stacked Generalization with Applicability Domain Outperforms Simple QSAR on in Vitro Toxicological Data. *J. Chem. Inf. Model.* **2019**, *59* (4), 1486–1496. https://doi.org/10.1021/acs.jcim.8b00553.

(17) Schaduangrat, N.; Homdee, N.; Shoombuatong, W. StackER: A Novel SMILES-Based Stacked Approach for the Accelerated and Efficient Discovery of ERα and ERβ Antagonists. *Sci. Rep.* **2023**, *13* (1), 21166. https://doi.org/10.1038/s41598-023-50393-w.

(18) Sheffield, T. Y.; Judson, R. S. Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Lethal Concentrations and Points of Departure. *Environ. Sci. Technol.* **2019**, *53* (21), 12793– 12802. https://doi.org/10.1021/acs.est.9b03957.

(19) Omotehinwa, T. O.; Oyewola, D. O.; Dada, E. G. A Light Gradient-Boosting Machine Algorithm with Tree-Structured Parzen Estimator for Breast Cancer Diagnosis. *Healthc. Anal.* **2023**, *4*, 100218. https://doi.org/10.1016/j.health.2023.100218.

(20) Sabherwal, G. Feature Selection Using Extra Trees Classifier for Parkinson's Disease Classification. *J. Mech. Cont. & Math. Sci.* **2024**, *Spl11* (1), 1– 12. https://doi.org/10.26782/jmcms.spl.11/2024.05.00010.

(21) Ghazwani, M.; Begum, M. Y. Computational Intelligence Modeling of Hyoscine Drug Solubility and Solvent Density in Supercritical Processing: Gradient Boosting, Extra Trees, and Random Forest Models. *Sci. Rep.* **2023**, *13* (1), 22492. https://doi.org/10.1038/s41598-023-37232-8.

(22) Bagheri, S.; Taridashti, S.; Farahani, H.; Watson, P.; Rezvani, E. Multilayer Perceptron Modeling for Social Dysfunction Prediction Based on General Health Factors in an Iranian Women Sample. *Front. Psychiatry* **2023**, *14*, 1283095. https://doi.org/10.3389/fpsyt.2023.1283095.

(23) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.

(24) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *Mol. Divers.* **2000**, *5* (4), 231– 243. https://doi.org/10.1023/a:1021372108686.

(25) Ferraroni, M.; Cornelio, B.; Sapi, J.; Supuran, C. T.; Scozzafava, A. Sulfonamide Carbonic Anhydrase Inhibitors: Zinc Coordination and Tail Effects Influence Inhibitory Efficacy and Selectivity for Different Isoforms. *Inorg. Chim. Acta* **2018**, *470*, 128–132. https://doi.org/10.1016/j.ica.2017.03.038.

(26) Supuran, C. T. Carbonic Anhydrases: Novel Therapeutic Applications for Inhibitors and Activators. *Nat. Rev. Drug Discov.* **2008**, *7* (2), 168–181. https://doi.org/10.1038/nrd2467.

(27) Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nat. Rev. Drug Discov.* **2014**, *13* (2), 105– 121. https://doi.org/10.1038/nrd4163.