Check for updates

Received: 20 December 2023 | Revised: 4 February 2024 | Accepted: 10 February 2024

DOI: 10.1002/jcc.27335

RESEARCH ARTICLE

Journal of COMPUTATIONAL CHEMISTRY WILEY

Explainable artificial intelligence in the design of selective carbonic anhydrase I-II inhibitors via molecular fingerprinting

Kevser Kübra Kırboğa[1,2] | Mesut Işık[1]

[1]Faculty of Engineering, Department of Bioengineering, Bilecik Seyh Edebali University, Bilecik, Turkey

[2]Bioengineering Department, Süleyman Demirel University, Isparta, Turkey

Correspondence
Kevser Kübra Kırboğa, Bilecik Seyh Edebali University, Department of Bioengineering, Faculty of Engineering, 11100, Bilecik, Turkey.
Email: kubra.kirboga@bilecik.edu.tr

Abstract

Inhibiting the enzymes carbonic anhydrase I (CA I) and carbonic anhydrase II (CA II) presents a potential avenue for addressing nervous system ailments such as glaucoma and Alzheimer's disease. Our study explored harnessing explainable artificial intelligence (XAI) to unveil the molecular traits inherent in CA I and CA II inhibitors. The PubChem molecular fingerprints of these inhibitors, sourced from the ChEMBL database, were subjected to detailed XAI analysis. The study encompassed training 10 regression models using IC_{50} values, and their efficacy was gauged using metrics including R^2, RMSE, and time taken. The Decision Tree Regressor algorithm emerged as the optimal performer (R^2: 0.93, RMSE: 0.43, time-taken: 0.07). Furthermore, the PFI method unveiled key molecular features for CA I inhibitors, notably PubChemFP432 (C(=O)N) and PubChemFP6978 (C(=O)O). The SHAP analysis highlighted the significance of attributes like PubChemFP539 (C(=O)NCC), PubChemFP601 (C(=O)OCC), and PubChemFP432 (C(=O)N) in CA I inhibitiotable n. Likewise, features for CA II inhibitors encompassed PubChemFP528(C(=O)OCCN), PubChemFP791 (C(=O)OCCC), PubChemFP696 (C(=O)OCCCC), PubChemFP335 (C(=O)NCCN), PubChemFP580 (C(=O)NCCCN), and PubChemFP180 (C(=O)NCCC), identified through SHAP analysis. The sulfonamide group (S), aromatic ring (A), and hydrogen bonding group (H) exert a substantial impact on CA I and CA II enzyme activities and IC_{50} values through the XAI approach. These insights into the CA I and CA II inhibitors are poised to guide future drug discovery efforts, serving as a beacon for innovative therapeutic interventions.

KEYWORDS
bioactivity, carbonic anhydrase, computational drug discovery, explainable artificial intelligence

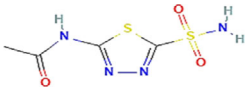
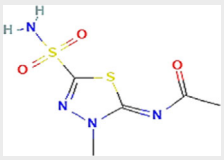
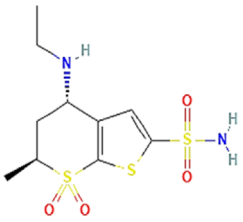
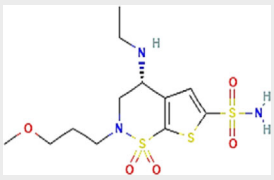
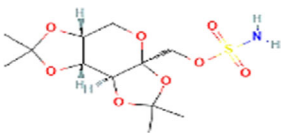
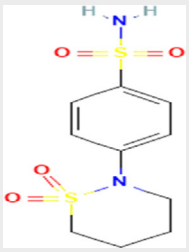
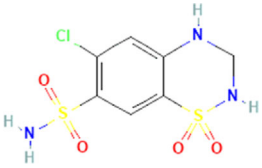
1 | INTRODUCTION

Carbonic anhydrase I and carbonic anhydrase II (CA I and CA II) act as catalysts for the hydration of carbon dioxide and hydrolysis of carbonic acid.[1,2] Red blood cells, in which CA I is highly expressed, are essential for transporting carbon dioxide and maintaining the pH balance. CA II is particularly highly expressed in neurons, eyes, and some glands. The role of CA II in neurons is to regulate the activation of neurons in response to pH changes.[3] The role of CA II in the eyes is the production and drainage of intraocular fluid and the regulation of intraocular pressure. In conclusion, the different structural properties and tissue distributions of CA I and CA II enable them to have different biological functions and play essential roles in regulating various physiological functions through the hydration of carbon dioxide and hydrolysis of carbonic acid.

CA I and CA II inhibitors affect the hydration of carbon dioxide and hydrolysis of carbonic acid by suppressing the catalytic activities of these enzymes.[4–6] Therefore, CA inhibitors are widely used in medical and biotechnological applications. Among the CA I inhibitors, acetazolamide is the most commonly used drug (Table 1). Acetazolamide treats glaucoma, high-altitude sickness, and epilepsy.[7] CA I inhibitors are also used to treat some types of cancer.[8,9]

 © 2024 Wiley Periodicals LLC.

TABLE 1 Current drugs are used for CA I and CA II inhibition.

CA I/C AII	Drug	Structure	2D
CA I CA II	Acetazolamide (diamox)	CC(=O)NC1 = NN=C(S1)S(=O)(=O)N	
CA I CA II	Methazolamide (neptazane)	CC(=O)N=C1N(N=C(S1)S(=O)(=O)N)C	
CA I CA II	Dorzolamide (trusopt)	CCNC1CC(S(=O)(=O)C2 = C1C=C(S2)S(=O)(=O)N)C	
CA I CA II	Brinzolamide (azopt)	CCNC1CN(S(=O)(=O)C2 = C1C=C(S2)S(=O)(=O)N)CCCOC	
CA I CAII	Topiramate (topamax)	CC1(OC2COC3(C(C2O1)OC(O3)(C)C)COS(=O)(=O)N)C	
CA I	Sultiame (ospolot)	C1CCS(=O)(=O)N(C1)C2 = CC=C(C=C2)S(=O)(=O)N	
CA I	Hydrochlorothiazide (HCTZ)	C1NC2 = CC(=C(C=C2S(=O)(=O)N1)S(=O)(=O)N)Cl	

Among the CA II inhibitors, dorzolamide is the most widely used drug. Dorzolamide is used to treat glaucoma and lowers intraocular pressure.[10] CA II inhibitors also treat central nervous system diseases such as cerebral ischemia and seizures.

CA I and CA II inhibitors are used in various research areas.[11] For example, CA inhibitors can also be used in applications such as biological water purification, carbon dioxide reduction, and biosensors.[12] CA inhibitors can also inhibit the growth of some microorganisms and are therefore also used as antimicrobial agents.[13] However, high CA I and CA II activity levels can cause various diseases. For example, increased CA I activity can lead to metabolic acidosis and impaired blood pH balance. Metabolic acidosis is characterized by decreased blood pH and can be seen in diseases such as kidney failure, diabetes, heart failure, and chronic obstructive pulmonary disease.[14,15] Increased CA II activity can cause increased intraocular pressure, leading to eye diseases such as glaucoma. In addition, excess CA II activity can cause cerebral ischemia, seizures, epilepsy, and other central nervous system diseases.[16] However, the complete absence or dysfunction of CA I

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

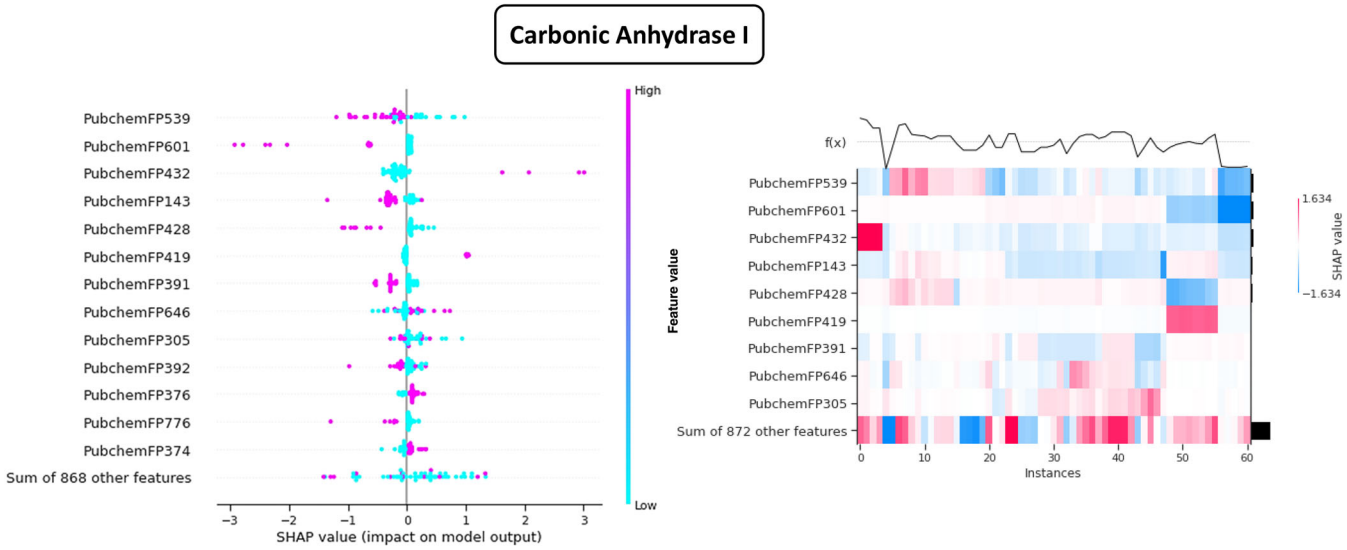


FIGURE 1 Shapley additive explanations (SHAP) results for CA I. Bee swarm and heatmap graph from left to right. The beeswarm graph presents the distrubiton of SHAP values for each feature. The heatmap shows the mean absolute value of SHAP values for each feature. This figure indicates which features are most sensitive to the CA I model. The most important features are marked in red on the heatmap. On the beeswarm graph, the points with high feature value are shown in neaon pink colour, and the points wiith kow feature importance value are shown in tuquoise colour. These colours visualise the magnitude of the contribution of the features to the model output.

and CA II inhibitors is rare. Typically, the body produces CA inhibitors, and these inhibitors are used to control enzyme activity. However, rare genetic disorders or drug side effects can affect CA inhibitors' production or functionality.[17] CA inhibitor deficiency can cause various diseases in these cases and should be treated with medications.

Drug discovery and design consist of long and complex steps such as in vivo and in vitro studies, cytotoxicity, and many bioactivity studies, as well as preclinical and clinical trials and manufacturing applications. All these stages are essential in designing an effective drug against a disease. However, the time spent developing many newly developed therapeutic agents, the increase in the number of employees, the difficulties experienced in the drug design and development process, and the increase in the anticipated costs are among the problems that negatively affect this process. Therefore, researchers turn to computational approaches such as virtual scanning (VS) and molecular insertion to minimize these problems, making effective drug designs before in vivo and in vitro studies. In addition, recently, new techniques such as artificial intelligence (AI), including deep learning (DL), and machine learning (ML) algorithms, have emerged as possible solutions to overcome the problems and obstacles in the drug design and discovery process.[18,19] Artificial intelligence methods can be used to discover CA inhibitors, and these technologies have great potential for the future discovery and development of CA inhibitors. This study estimated whether or not the compounds obtained by virtual scanning of CA I and CA II inhibitors in the ChEMBL database contain molecular fingerprints by explicable artificial intelligence (XAI) methods over IC_{50} values[20] (Figure 1).

This study analyzed the inhibitory effects and IC_{50} values of a new molecule on CA I and CA II enzymes and existing drugs with the XAI method. Our study aims to reveal the impact of small bit structures in molecular fingerprints of inhibitors in existing databases on CA I and CA II enzyme activities and IC_{50} values. Our primary research question is whether the XAI method can determine the importance of these inhibitors, which have a role in many multisystem diseases. For this purpose, we developed a model algorithm according to the presence or absence of molecular fingerprints of inhibitors obtained from the ChEMBL database. By integrating this algorithm with XAI, we highlighted which small bit structures are more important. This study, which makes up for the lack of use of XAI on CA I and CA II enzymes in the literature, will guide the design of future CA I and CA II enzyme inhibitors.

This study consists of five parts. The introduction introduces the study's subject, background, aim, and research question. The method section explains data set collection, preprocessing, creation of machine learning algorithms, and descriptive artificial intelligence (XAI) integration. The findings section presents the evaluation of algorithms with different performance criteria and the visualization of the results obtained by the XAI method. In the discussion section, the comparison of the findings with the literature and the contribution and importance of the study were evaluated. In the conclusion part, the limitations of the study, its advantages, and its impact on future studies are stated.

2 | RESULTS AND DISCUSSION

2.1 | Permutation feature importance

The results obtained using the permutation feature importance (PFI) method determined the order of importance of molecular fingerprint features for CA I and CA II enzymes. For CA I, PubChemFP432, PubChemFP528, PubChemFP336, PubChemFP391, PubChemFP286, and PubChemFP366 were identified as the essential features (Table 2). The most important CA II features were PubChemFP528, PubChemFP374, PubChemFP593, PubChemFP647, and PubChemFP715 (Table 2). These

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Molecular fingerprint features found with the permutation feature importance method for CA I and CA II.

	Permutation feature importance score	Feature	Bit substructure
CA I	0.1911 ± 0.1187	PubchemFP432	C(—C)(—C)(═O)
	0.0439 ± 0.0396	PubchemFP528	[#1]—N—C—[#1]
	0.0326 ± 0.0288	PubchemFP336	C(∼C)(∼C)(∼C)(∼N)
	0.0256 ± 0.0291	PubchemFP391	N(∼C)(∼C)(∼C)
	0.0234 ± 0.0200	PubchemFP286	C—O
	0.0188 ± 0.0531	PubchemFP366	C(∼H)(∼O)
	0.0182 ± 0.0141	PubchemFP643	[#1]—C—C—N—[#1]
	0.0143 ± 0.0288	PubchemFP586	C—S—C—C—C
	0.0121 ± 0.0223	PubchemFP338	C(∼C)(∼C)(∼H)(∼N)
CA II	0.0560 ± 0.0152	PubchemFP528	[#1]—N—C—[#1]
	0.0153 ± 0.0211	PubchemFP374	C(∼H)(∼H)(∼H)
	0.0070 ± 0.0107	PubchemFP593	N—C—C—C—N
	0.0070 ± 0.0072	PubchemFP647	O═C—N—C—N
	0.0057 ± 0.0022	PubchemFP715	Cc1ccc(S)cc1
	0.0053 ± 0.0007	PubchemFP17	> = 8 N
	0.0042 ± 0.0062	PubchemFP258	> = 2 hetero-aromatic rings
	0.0041 ± 0.0022	PubchemFP778	CC1CCC(S)CC1
	0.0037 ± 0.0058	PubchemFP643	[#1]—C—C—N—[#1]
	0.0036 ± 0.0011	PubchemFP356	C(∼C)(:C)(:C)

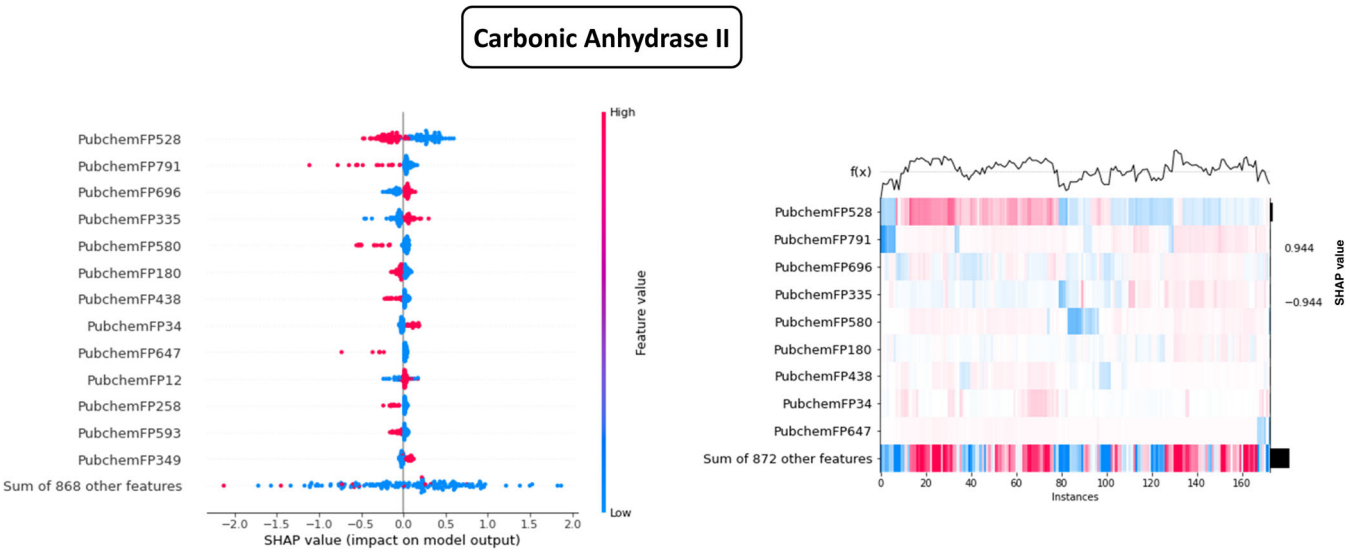


FIGURE 2 Shapley additive explanations (SHAP) results for CA II. Bee swarm and heatmap graph from left to right. The bee swarm graph displays the distribution of SHAP values for each feature. The heatmap graph illustrates the mean absolute value of SHAP values for each feature. This figure reveals which features are most sensitive to the CA II model. The most important features are marked in red on the heatmap. On the bee swarm graph, the points with high feature value are shown in red color, and the points with low feature value are shown in blue color.

results show that molecular fingerprint features have different ordering of importance for CA I and CA II enzymes.

2.2 | Model development and validation

Among the 10 models trained for CA I and CA II inhibitors, the decision tree regression algorithm was selected as the best-performing model with 0.93 R^2, 0.43 RMSE, and 0.07 time-taken values.

2.3 | Shapley additive explanations

According to the results of the SHapley additive exPlanations (SHAP) method, which explains according to Shapley values, which is one of the XAI methods, PubchemFP539, PubchemFP601, PubchemFP432, PubchemFP143, PubchemFP428 molecular fingerprints are of great importance for CA I (Figure 1), while PubchemFP528, PubchemFP791, PubchemFP696, PubchemFP335, PubchemFP580, and PubchemFP180 fingerprints are of great significance to CA II (Figure 2).

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

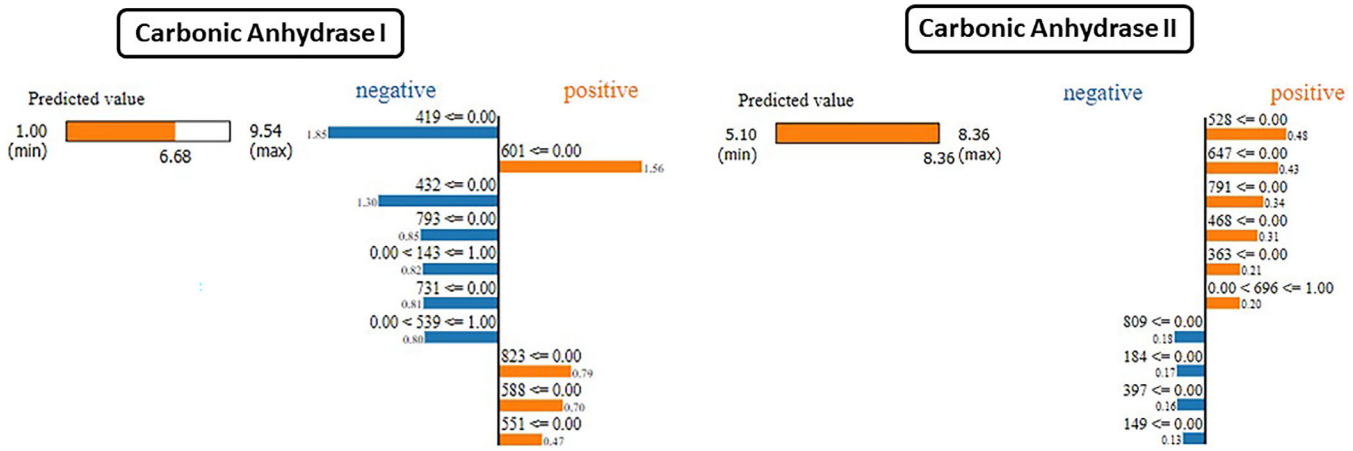


FIGURE 3 Local interpretable model-agnostic explanations (LIME) explainability results for CA I and CA II. LIME shows the explainability results of the CA I model for a data sample with a predicted value of 6.68. The maximum value of the CA I sample is 9.54 and the minimum value is 1.00. This means that the data sample is predicted to be moderately active by the CA I model. In the CA II example, it means that the maximum value is actively estimated (max 8.36). LIME explains what features these predictions are based on and how important these features are. In this way, LIME is a useful tool to support the decision-making process.

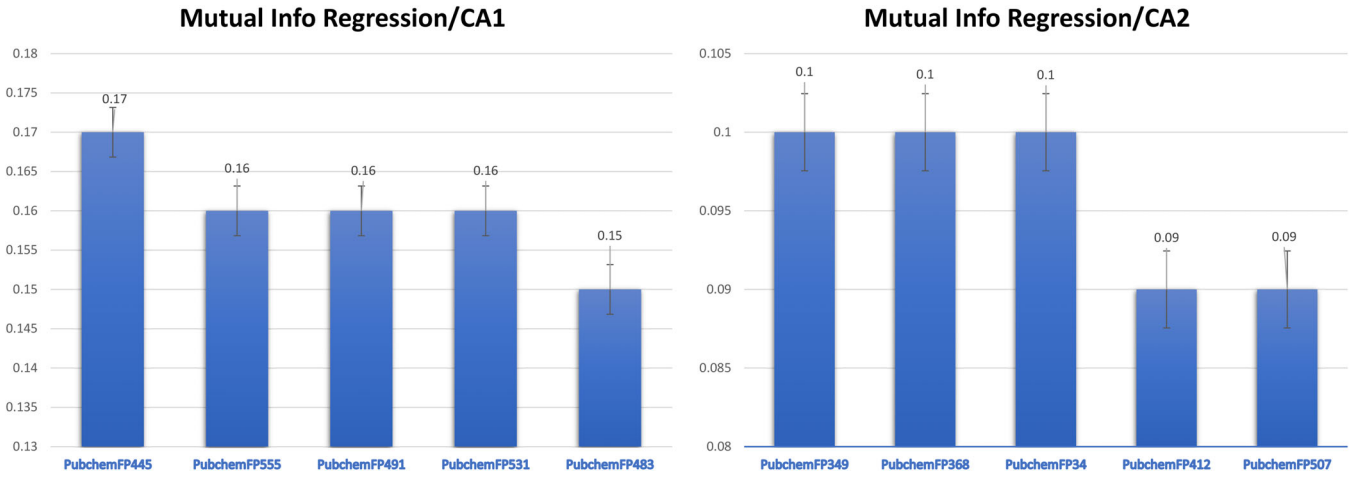


FIGURE 4 Mutual information regression (MIR) results for CA I and CA II.

2.4 | Local interpretable model-agnostic explanations

Virtual scanning of CA I and CA II inhibitors and whether they contain molecular fingerprints were estimated by XAI methods. The local interpretable model-agnostic explanations (LIME) algorithm has revealed the accuracy of these predictions. While CA I has an accuracy rate of about 70%, CA II has an accuracy of 100% (Figure 3).

2.5 | Mutual information regression

According to the results of the mutual information regression (MIR) analysis, PubchemFP445, Pubchem555, and PubchemFP491 contributed the most to the output variable for CA I, while PubchemFP349, PubchemFP368, PubchemFP34 contributed significantly to the output variable for CA II (Figure 4).

According to XAI and standard feature importance methods, PubchemFP432 was most important for CA I, while PubchemFP528 was for CA II. In Table 3, the bit substructure structures of these features according to the bit positions are shown in detail.

Molecular fingerprints are a set of numerical values used to describe the structure of a molecule. These values represent various features in the molecule's structure and measure similarities and differences between molecules. Molecular fingerprint similarity measures structural similarity between two or more molecules. This similarity may be due to molecules containing the same or similar chemical groups, bonding of atoms similarly, or having similar physical properties. Interpretation of molecular fingerprint similarity can vary in many different contexts. For example, compounds with similar

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Top five-bit substructures according to their significant scores according to bit positions in CA I and CA II inhibitors.

Methods	CA I		CA II	
PFI	PubchemFP432	C(—C)(—C)(═O)	PubchemFP528	[#1]—N—C—[#1]
	PubchemFP528	[#1]—N—C—[#1]	PubchemFP374	C(∼H)(∼H)(∼H)
	PubchemFP336	C(∼C)(∼C)(∼C)(∼N)	PubchemFP593	N—C—C—C—N
	PubchemFP391	N(∼C)(∼C)(∼C)	PubchemFP647	O═C—N—C—N
	PubchemFP286	C—O	PubchemFP715	Cc1ccc(S)cc1
Explainable AI	PubchemFP539	N═C—C—[#1]	PubchemFP528	[#1]—N—C—[#1]
	PubchemFP601	N—C:C: C—N	PubchemFP791	NC1CCC(N)CC1
	PubchemFP432	C(—C)(—C)(═O)	PubchemFP696	C—C—C—C—C—C—C—C
	PubchemFP143	> = 1 any ring size 5	PubchemFP335	C(∼C)(∼C)(∼C)(∼H)
	PubchemFP428	C(#C)(—H)	PubchemFP580	O═C—C—C—N
MIR	PubchemFP445	C(—H)(—N)(═C)	PubchemFP349	C(∼C)(∼H)(∼S)
	PubchemFP555	N═C—C═C	PubchemFP368	C(∼H)(∼S)
	PubchemFP491	C—N:C—[#1]	PubchemFP34	O—P
	PubchemFP531	S—C:C—C	PubchemFP412	S(∼C)(∼C)
	PubchemFP483	N—S—C:C	PubchemFP507	C—C—S—C

molecular fingerprints may exhibit similar biological activities in a drug design project. Therefore, molecular fingerprint similarity can be used to predict the biological activity of a compound.[21]

On the other hand, molecular fingerprint similarity can also be used to estimate the strength of chemical or physical interactions between two molecules. Molecules with similar molecular fingerprints tend to have stronger interactions with each other. In summary, molecular fingerprint similarity is a powerful tool for measuring molecule similarities and differences. Interpreting this similarity can vary in different contexts and can be used to predict molecules' chemical or biological properties.[22,23]

Permutation importance (PFI) and mutual info regression (MIR) are common methods for generating feature importance ranking. The PFI method measures the impact of random changes in a feature value on model performance. However, this method does not consider interactions between features and only measures the effect of each feature on model performance. Conversely, MIR determines the order of importance of the features based on the features' mutual information values of the features in the target variable. However, this method also does not consider interactions between features and, in some cases, can cause computational difficulties in high-dimensional data sets.[24,25]

On the other hand, the LIME method is developed to make a local explanation for each prediction. This method is based on a model that mimics the operation of the model on a given sample and measures its contribution to the model prediction for each feature.[26] LIME does not account for interactions like SHAP but can be faster and less computationally intensive on smaller-sized data sets. As a result, the SHAP and LIME methods are better options for generating feature importance rankings because they can account for interactions between features and work effectively even on high-dimensional data sets. The SHAP and LIME methods are feature importance ranking methods developed to correct these disadvantages. The SHAP method measures the contribution of each attribute to the model prediction using a mathematical concept called Shapley values. This method considers interactions and can also work effectively on multidimensional data sets.[27]

In this study, the effect range and distribution of the results obtained according to the SHAP method are presented in Figures 3 and 4, where each line is a fingerprint, and each point is a sample. The x-axis represents the SHAP values that can identify an effect that increases positively and decreases negatively. If the PubchemFP539 feature is present in a molecule, it has a negative SHAP value (purple); if not, it has a positive SHAP value (blue). The PubchemFP539 trait may cause a high variation in the estimated IC_{50} values. PubchemFP539 may be useful in inhibitor design and discovery for the CA I enzyme, as the decrease in IC_{50} represents a high inhibitory potential. Likewise, PubchemFP601, PubchemFP432, PubchemFP143 and PubchemFP428 for CA I and PubchemFP528, PubchemFP791, PubchemFP696, PubchemFP335, PubchemFP580, and PubchemFP180 fingerprints for CA II can improve inhibitory activity. According to the PFI and Explainable AI method, PubchemFP432 for CA I and PubchemFP528 bit substructures for CA II can improve inhibitory activity (Table 3). Characteristics with high values could contribute more to the estimation of inhibitor potential.

If fingerprints important to CA I and II are present in the compounds, they tend to increase the efficacy of potential inhibitors. The structures of acetazolamide and methazolamide compounds are used both as active pharmaceutical ingredients in treating diseases (Table 1) and as standard reference compounds for these enzymes due to their inhibition potential against CA I and II, as well as subgroups are presented in Table 4. The studies emphasized that most of the bit substructures that make up the molecular fingerprint are the components of the basic skeletons of the compounds used as standard inhibitors.[28] In this study, bit substructures such as PubchemFP432 (carbonyl group), PubchemFP491, PubchemFP528 (amine group), PubchemFP483 (sulfur atom and amine group), PubchemFP593 (nitrogen-containing heterocyclic rings) and

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Substructures corresponding to fingerprints with significant impact on CA I and II.

Standard inhibitors (drug)	Bit substructure				
Acetazolamide	PubchemFP432	PubchemFP491	PubchemFP528	PubchemFP593	PubchemFP647
Methazolamide	PubchemFP483				

PubchemFP647 (carbonyl and amine group combination), which are effective on potential inhibitory structures as well as the molecular structure of acetazolamide and methazolamide are presented in Table 3. The bit substructures showed similarities with many groups in the basic skeletons of the reference compounds (acetazolamide and methazolamide). These similarities are presented by marking the groups with the same color (Table 3). Finding these similarities between the bit substructures and standard compounds may indicate that a suitable method was used to estimate the inhibitor potential.

Compounds such as dichlorophenamide, dorzolamide, ethoxolamide, acetazolamide, and methazolamide, which are widely used for treatment in the clinic, are among the sulfonamide drugs.[29,30] In addition, sulfonylamide group compounds are widely used as potential inhibitors in treatments based on the inhibition of discovered human carbonic anhydrase isozymes. The literature studies tested the inhibitory potentials of clinically used derivatives such as acetazolamide, methazolamide, dorzolamide, ethoxolamide, and dichlorophenamide due to their interactions with CA. While many sulfanilamide derivatives have a CA inhibitory potential at micromolar (K_I in the range of 1.46–6.50 μM), derivatives such as acetazolamide/metazolamide showed an inhibitory effect in the range of 15–48 nM.[30] Therefore, in study results, we observed that sulfonamide inhibitors performed better than other classes. This proves that the algorithms we used in our study play an important role in the design of inhibitors for CA I and CA II.

CA inhibitors are used in different medical applications by blocking the activity of CA enzymes. CA I and CA II inhibitors can be of many different natures. However, some molecules have common structures for similar enzymes, such as CA I and CA II. These common structures can enable inhibitors to bind to CA enzymes and act as inhibitors. For example, sulfonamides are one of CA inhibitors' most widely used classes. These molecules have a structure similar to the active site of CA enzymes and form the inhibition effect by binding strongly with the enzyme.

Similarly, some thiadiazole and thiadiazoline compounds can act as CA I and CA II inhibitors. In many studies conducted in recent years, high inhibition effects of sulfanilamide derivatives against CA I and CA II and important subgroups that increase the inhibition efficiency have been determined.[30–34] Therefore, it can be stated that the bit substructures determined by PFI, Explainable AI, and MRI methods may show similarities with many groups in the basic skeletons of sulfonamide derivative compounds in the literature studies.

These results will explain the restructuring-activity relationships for CA I and II inhibitors and shed light on discovering and designing new inhibitors with effective inhibitory potential. Thus, very time-consuming bioactive studies will be carried out quickly, reducing costs.

3 | EXPERIMENTAL

3.1 | Data set

This study collected bioactivity data for human CA I and CA II from version 25 of the ChEMBL database. ChEMBL is a hand-curated

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

database of bioactive molecules with drug-like properties, combining chemical, bioactivity and genomic data to help translate genetic information into effective new drugs.[35] For CA I and CA II, IC_{50} values were chosen as the unit of bioactivity. IC_{50} is a measure of the potency of a substance in inhibiting a specific biological or biochemical function. IC_{50} selection was accomplished through a data curation process. The data curation process is a method used to detect and correct missing, inconsistent, or abnormal values in a data set. In this process, IC_{50} values for CA I and CA II were filtered by removing values 3 standard deviations away from the data set's mean. Thus, a data set of 372 compounds for CA I and 865 compounds for CA II was obtained. Since this study aimed to establish a regression model for CA I and CA II, threshold values were determined to distinguish active and inactive compounds. These threshold values were chosen as <1 and >10 μM, commonly used in the literature. According to these criteria, 67 active and 305 inactive compounds were classified for CA I and 140 active and 725 inactive compounds for CA II. These composites were input to the regression model.[15,36,37]

3.2 | Molecular descriptor

Molecular fingerprint identifiers for the compounds in the obtained data sets were calculated using the PaDEL-Descriptor software.[22] SMILES indicators were used to calculate these molecular descriptors. Compound structures have been standardized using functions included in the PaDEL software. Molecular fingerprints play a crucial role in QSAR studies as they identify molecules and characterize chemical structure information quantitatively and qualitatively.[22,38]

3.3 | Data splitting

We obtained a data collection that ML models can handle and split into two subgroups using random sampling. This method ensures equal data distribution in each subgroup while preserving the homogeneity of the data set.[39] The training set has a larger data percentage (80%) and is used to develop the model. It contains 298 samples for CA I and 692 samples for CA II. The test set has a more limited data percentage (20%) and is used to test the model. It contains 74 samples for CA I and 173 samples for CA II. Both subgroups have similar class ratios and statistical properties for weight and pitch variables, indicating that the data set is well split and the model is generalizable.

The construction of mathematical descriptors frequently involves a large number of variables. The training set, on the other hand, looks for the best variable subset that has the right and required data. The number of extraneous variables is minimized in this method. A subgroup of characteristics is added to the initial set without altering the variables' contents to give a biologically plausible explanation.[40] The model is trained following the selection of the ideal subset of variables. To maintain the model's validity while dealing with ambiguous data, overtraining should be avoided.

In these circumstances, cross-validation (CV) methods are frequently employed. In CV, performance evaluation of the model, performance prediction for unknown data, and generalization degree measurement are all included. At the beginning of each iteration of the experiment in the CV, the original data set is divided into two subgroups (training set and validation set). In our investigation, the 10-fold CV method was employed. The goal of a CV is to give you the tools necessary to choose the perfect collection of criteria. These parameters evaluate each model's performance, and the model with the best performance is selected. The final validation of the best model is then completed. If the validation outcomes are statistically significant, it may be said that a unique predictive pharmacological model has been constructed.[41–43]

3.4 | Permutation feature importance

PFI is a method that evaluates how much each feature or variable in a machine learning model contributes to the model's performance.[24] This method generates a random permutation on each feature, measuring how much a decrease in the model's performance is caused. This method first applies an appropriate machine learning model to the model using the training data set. Then, random permutations are generated on each feature, and the model is retrained using these permutations. Next, the performance of the retrained model is recorded. Finally, the performance loss caused by the permutations for each feature is calculated, and the ranking of the features is determined according to these losses. The more performance loss a feature causes, the higher its importance is considered. In this study, the PFI method in the molecular fingerprint identification phase accompanying the virtual scanning process of CA I and CA II was performed using the permutation_importance() function of the sci-kit-learn library.

3.5 | Model development and validation

With the target data sets taken from the ChEMBL database, the 'train_test_split' function is divided into 80% training and 20% test set. Classes and functions are provided for machine learning algorithms using the sci-kit-learn library. Ten model algorithms were used, including RandomForestRegressor, MLPRegressor, XGBRegressor, GaussianProcessRegressor, and DecisionTreeRegressor. To evaluate the models on the test set, the test set data were fed with the model, and Adjusted *R*-squared, *R*-squared, RMSE, and Time Taken parameters were used to evaluate the model's performance. The results were visualized with the Seaborn v0.8.1 Python package.[44]

3.6 | SHapley additive exPlanations

SHAP is a method used to describe the predictions of machine learning models. This method attempts to explain the contribution of each

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

feature and its effect on the prediction. In this way, it can be used to understand why models make a prediction. SHAP calculates the contributions for each attribute using a concept called Shapley values.[27,45] Shapley values are a game theory concept based on the distribution of the gains of a coalition of players to each player in the alliance.[46] If we consider each attribute as a player, combinations of attributes can also be regarded as coalitions. In this way, an attribute receives a Shapley value based on its contribution within all coalitions. SHAP estimates the contribution of each attribute using Shapley values. The contributions of the attributes differ according to the attribute values in each sample. Therefore, SHAP calculates estimates using the attribute values of each sample and, as a result, returns a SHAP value that determines the contribution of each attribute in a sample. SHAP can be used for many machine learning models, especially black box models. It is often used to increase explainability. The SHAP library provides many functions that can be used in Python. This study used a tree-based "shap.TreeExplainer" function among SHAP functions.

3.7 | Local interpretable model-agnostic explanations

LIME is a method used to describe the predictions of machine learning models. This method creates a local model to explain the predictions in each sample and tries to explain each feature contribution based on this model.[47] LIME makes a local model to calculate contributions for each attribute. This local model is optimized to mimic the primary model's decision best. This way, it determines the contributions of the attributes in each sample and, as a result, returns a LIME value that determines the effect of each attribute. The lime.lime_tabular function was used in this study. This function takes a sample data set and a prediction model and returns a LIME value that explicitly determines each attribute's contribution to a sample.

3.8 | Mutual information regression

MIR is a feature selection method. This method measures the relationship of the features in the data with the target variable and selects the best features based on this relationship. MIR uses the concept of mutual information (MI) to measure the information exchange between two variables. MI is a measure that measures the relationship between two variables.[48] The higher the relationship between the two variables, the higher the MI value. MIR calculates the MI with the target variable for each feature. Then, the features with the highest MI value are selected. These features have the most substantial relationship with the target variable and can be used to improve the model's performance. This study implemented the MIR method using the mutual_info_regression function in the sklearn.feature_selection module in Python. This function calculated the MI values between each attribute in the data set and the target variable.

4 | CONCLUSION

The study selected compounds with high inhibition potential against CA I and CA II enzymes. The bit substructure of compounds with effective inhibitory potential was determined using the XAI approach. Our research questions the ability of the XAI method to determine the importance of these inhibitors in the context of multisystem diseases. The algorithmic model we developed classifies inhibitors in the ChEMBL database according to the presence or absence of molecular signatures. Thanks to the integration with XAI, the small structural units with the most significant impact on the activities and IC_{50} values of CA I and CA II enzymes were successfully highlighted. In particular, these structural units that show metal ion binding, hydrophobic interactions, and polarity-related properties provide essential clues to increasing the therapeutic potential of inhibitors. However, our study also has some limitations; these include using only inhibitors derived from a single database, investigating only CA I and CA II enzymes, and evaluating small structural units alone. Future studies should include inhibitors obtained from various databases, a broader range of CA enzyme types, and various molecular signatures. This study pioneered the application of XAI in the study of CA I-II enzymes, filling an essential gap in the existing literature and providing valuable insights into developing inhibitors targeting CA I-II enzymes.

AUTHOR CONTRIBUTIONS

Mesut Işık: Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); validation (equal); visualization (equal); writing—original draft (equal). **Kevser Kübra Kırboğa:** Conceptualization (equal); funding acquisition (equal); software (equal); resources (equal); writing—review and editing (equal).

ACKNOWLEDGEMENTS

We would like to thank Prof. Dr. Ecir Uğur Küçüksille for his invaluable guidance and support throughout our research. He provided us with constructive feedback, insightful suggestions, and useful resources that helped us improve the quality of our work. We are grateful for his patience, encouragement, and expertise.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in CHEMBL at https://www.ebi.ac.uk/chembl/, Reference 35.

ORCID

Kevser Kübra Kırboğa https://orcid.org/0000-0002-2917-8860
Mesut Işık https://orcid.org/0000-0002-4677-8104

REFERENCES

[1] S. Bekku, H. Mochizuki, E. Takayama, N. Shinomiya, H. Fukamachi, M. Ichinose, T. Tadakuma, T. Yamamoto, *Res Exp Med* **1998**, *198*, 175.

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[2] S. Talekar, B. H. Jo, J. S. Dordick, J. Kim, *Curr. Opin. Biotechnol.* **2022**, *74*, 230.
[3] S. Pastorekova, S. Parkkila, J. Pastorek, C. T. Supuran, *J. Enzyme Inhib. Med. Chem.* **2004**, *19*, 199.
[4] C. T. Supuran, A. Scozzafava, A. Casini, *Med. Res. Rev.* **2003**, *23*, 146.
[5] C. T. Supuran, D. Vullo, G. Manole, A. Casini, A. Scozzafava, *Curr. Med. Chem. Cardiovasc. Hematol. Agents* **2004**, *2*, 51.
[6] C. T. Supuran, A. S. A. Altamimi, F. Carta, *Expert Opin. Ther. Pat.* **2019**, *29*, 781.
[7] S. Mussi, S. Rezzola, P. Chiodelli, A. Nocentini, C. T. Supuran, R. Ronca, *J. Enzyme Inhib. Med. Chem.* **2022**, *37*, 280.
[8] A. Mentese, E. Fidan, A. Alver, S. Demir, S. O. Yaman, A. Sumer, S. Fidan, H. Kavgaci, I. Turan, *Cent Eur J Immunol* **2017**, *42*, 73.
[9] C. T. Supuran, *Expert Opin. Investig. Drugs* **2018**, *27*, 963.
[10] K. Boyne, D. A. Corey, P. Zhao, B. Lu, W. F. Boron, F. J. Moss, T. J. Kelley, *Am. J. Physiol. Lung Cell. Mol. Physiol.* **2022**, *322*, L333.
[11] P. A. Preisig, R. D. Toto, R. J. Alpern, *Ren. Physiol.* **1987**, *10*, 136.
[12] C. T. Supuran, *Mar. Drugs* **2022**, *20*, 721.
[13] C. T. Supuran, *J. Enzyme Inhib. Med. Chem.* **2021**, *36*, 1702.
[14] M. Adamczak, A. Masajtis-Zagajewska, O. Mazanowska, K. Madziarska, T. Stompór, A. Więcek, *Kidney Blood Press. Res.* **2018**, *43*, 959.
[15] H. J. Kim, *Electrolyte Blood Press* **2021**, *19*, 29.
[16] L. Ciccone, C. Cerri, S. Nencetti, E. Orlandini, *Molecules* **2021**, *26*. https://doi.org/10.3390/molecules26216380
[17] S. Lahiri, A. Roy, S. M. Baby, T. Hoshi, G. L. Semenza, N. R. Prabhakar, *Prog. Biophys. Mol. Biol.* **2006**, *91*, 249.
[18] W. Duch, K. Swaminathan, J. Meller, *Curr. Pharm. Des.* **2007**, *13*, 1497.
[19] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, P. Kumar, *Mol. Divers.* **2021**, *25*, 1315.
[20] K. K. Kırboğa, E. U. Küçüksille, M. E. Naldan, M. Işık, O. Gülcü, E. Aksakal, *Comput. Methods Prog. Biomed.* **2023**, *233*, 107492.
[21] L. Xue, J. Bajorath, *Comb. Chem. High Throughput Screen.* **2000**, *3*, 363.
[22] C. W. Yap, *J. Comput. Chem.* **2011**, *32*, 1466.
[23] D. Probst, J.-L. Reymond, *J. Chem.* **2018**, *10*, 1.
[24] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, *Bioinformatics* **2010**, *26*, 1340.
[25] M. Bennasar, Y. Hicks, R. Setchi, *Expert Syst. Appl.* **2015**, *42*, 8520.
[26] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, E. Costa da Silva, *Sensors* **2019**, *19*, 2969.
[27] J. Derks, H. Peters, *Int. J. Game Theory* **1993**, *21*, 351.
[28] Y. Wu, M. Li, J. Shen, X. Pu, Y. Guo, *Mol. Divers.* **2023**. https://doi.org/10.1007/s11030-023-10649-z
[29] J. G. Stock, *Arch. Ophthalmol.* **1990**, *108*, 634.
[30] I. Nishimori, D. Vullo, A. Innocenti, A. Scozzafava, A. Mastrolorenzo, C. T. Supuran, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3828.
[31] N. Lolak, S. Akocak, M. Durgun, H. E. Duran, A. Necip, C. Türkeş, M. Işık, Ş. Beydemir, *Mol. Divers.* **2022**, *27*, 1735.
[32] M. Işık, S. Akocak, N. Lolak, P. Taslimi, C. Türkeş, İ. Gülçin, M. Durgun, Ş. Beydemir, *Arch Pharm* **2020**, *353*, e2000102.
[33] M. Durgun, C. Türkeş, M. Işık, Y. Demir, A. Saklı, A. Kuru, A. Güzel, Ş. Beydemir, S. Akocak, S. M. Osman, Z. AlOthman, C. T. Supuran, *J. Enzyme Inhib. Med. Chem.* **2020**, *35*, 950.
[34] D. Vullo, J. Voipio, A. Innocenti, C. Rivera, H. Ranki, A. Scozzafava, K. Kaila, C. T. Supuran, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 971.
[35] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, et al., *Nucleic Acids Res.* **2017**, *45*, D945.
[36] S. Lamotte, N. Aulner, G. F. Späth, E. Prina, *Sci. Rep.* **2019**, *9*, 438.
[37] S. Chiba, K. Ikeda, T. Ishida, M. M. Gromiha, Y. h. Taguchi, M. Iwadate, H. Umeyama, K.-Y. Hsin, H. Kitano, K. Yamamoto, et al., *Sci. Rep.* **2015**, *5*, 17209.
[38] A. A. Malik, C. Phanus-Umporn, N. Schaduangrat, W. Shoombuatong, C. Isarankura-Na-Ayudhya, C. Nantasenamat, *J. Comput. Chem.* **2020**, *41*, 1820.
[39] R. Pramoditha, *Why Do We Set A Random State in Machine Learning Models?* Vol. 2024, Towards Data Science, **2022**.
[40] Y. Saeys, I. Inza, P. Larranaga, *Bioinformatics* **2007**, *23*, 2507.
[41] P. Carracedo, J. Blanco, N. Rodriguez-Fernandez, F. Cedrón-Santaeufemia, F. Nóvoa, A. Carballal, V. Maojo, A. Pazos, C. Fernandez-Lozano, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538.
[42] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, et al., *J. Med. Chem.* **2014**, *57*, 4977.
[43] P. Gramatica, A. Sangion, *J. Chem. Inf. Model.* **2016**, *56*, 1127.
[44] M. Waskom, O. Botvinnik, D. Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, A. Qalieh, mwaskom/seaborn: v0.8.1. *Zenodo*. **2017** https://doi.org/10.5281/zenodo.883859
[45] A. Joseph, *Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models*, Bank of England, London **2019**. https://doi.org/10.48550/arXiv.1903.04209
[46] A. E. Roth, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, Cambridge University Press, Cambridge **1988**, p. 1.
[47] A. Junkang, Y. Zhang, and I. Joe, Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models. *Appl. Sci.* **2023**, *13*, 8782. https://doi.org/10.3390/app13158782
[48] Guhanesvar, Feature Selection Based on Mutual Information Gain for Classification and Regression. **2021**.

How to cite this article: K. K. Kırboğa, M. Işık, *J. Comput. Chem.* **2024**, *45*(18), 1530. https://doi.org/10.1002/jcc.27335

1096987x, 2024, 18, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27335 by Bilecik Seyh Edebali, Wiley Online Library on [05/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License