Computational and Structural Biotechnology Journal 23 (2024) 3669–3679

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Rapid discovery of Transglutaminase 2 inhibitors for celiac disease with boosting ensemble machine learning

Ibrahim Wichka [1], Pin-Kuang Lai [*,2]

Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Keywords:
Computational drug discovery
Ensemble machine learning
Inhibitor screening
Transglutaminase 2
Celiac disease
Quantitative structure-activity relationship (QSAR)

ABSTRACT

Celiac disease poses a significant health challenge for individuals consuming gluten-containing foods. While the availability of gluten-free products has increased, there is still a need for therapeutic treatments. The advancement of computational drug design, particularly using bio-cheminformatics-oriented machine learning, offers promising avenues for developing such therapies. One promising target is Transglutaminase 2 (TG2), a protein involved in the autoimmune response triggered by gluten consumption. In this study, we utilized data from approximately 1100 TG2 inhibition assays to develop ligand-based molecular screening techniques using ensemble machine-learning models and extensive molecular feature libraries. Various classifiers, including tree-based methods, artificial neural networks, and graph neural networks, were evaluated to identify primary systems for predictive analysis and feature significance assessment. Boosting ensembles of perceptron deep learning and low-depth random forest weak learners emerged as the most effective, achieving over 90 % accuracy, significantly outperforming a baseline of 64 %. Key features, such as the presence of a terminal Michael acceptor group and a sulfonamide group, were identified as important for activity. Additionally, a regression model was created to rank active compounds. We developed a web application, Celiac Informatics (https://celiac-informatics-v1–2b0a85e75868.herokuapp.com), to facilitate the screening of potential therapeutic molecules for celiac disease. The web app also provides drug-likeness reports, supporting the development of novel drugs.

1. Introduction

Celiac disease is a prominent gluten-related disease that is prevalent in about 1 % of the world's population [1]. However, the pathogenesis and the role of a target protein in these diseases were not understood until advanced in vitro methods for target discovery and cell culture analysis emerged [2]. Samuel Gee constructed the first clinical description of celiac disease in 1887, detailed dietary treatment, and described the best method of managing patient symptoms. Willem Carel Dicke [3] developed the first formal wheat-free diet in the 1940s [4]. Thereafter, the focus has shifted to gluten as a key factor in pathogenesis, leading to diets excluding wheat, barley, and rye. The gluten-free diet (GFD) became the main treatment for celiac patients, but its long-term sustainability is questioned. Strict adherence requires vigilance, and the prevalence of gluten in essential foods makes it even more challenging. Many of those on GFDs experience a decline in well-being, with two-thirds not fully recovering. 80 % of GFD patients also end up mistakenly consuming gluten when on the diet and the diet itself has

Abbreviations: TG2, Transglutaminase 2; TG3, Transglutaminase 3; GFD, Gluten-free diet; HLA-DQ2, Human leukocyte antigen DQ2; QSAR, Quantitative structure-activity relationship; LBDD, Ligand-based drug design; IC50, Half-maximal inhibitory concentration; SMILES, Simplified molecular input line entry system; AdaBoost, Adaptive Boosting; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; MLP, Multi-layer-perceptron; NumSaturatedRings, Number of saturated rings; NumSaturatedHeteroycles, Number of saturated heterocycles; SMR_VSA6, substitution matrix representation of van der Waals surface area (VSA); VSA_Estate3, sum of VSA values for a certain E-state (Electrotopological state); BCUT2D_CHGHI, highest BCUT (derived from Burden matrix) value considering atomic charges; SlogP_VSA2, highest BCUT (derived from Burden matrix) value considering atomic charges.; BCUT2D_CHGLO, lowest BCUT (derived from Burden matrix) value considering atomic charges; PEOE_VSA12, partial Equalization of Orbital Electronegativities effect on Van der Waals Surface Area; BCUT2D_MRHI, highest BCUT value considering molecular refractivity (MR), representing the polarizability of molecule; PEOE_VSA8, partial Equalization of Orbital Electronegativities effect on Van der Waals Surface Area.; SVM, Support Vector Machine.

* Corresponding author at: Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA.
E-mail address: plai3@stevens.edu (P.-K. Lai).
[1] ORCID: 0009-0009-6380-5272
[2] ORCID: 0000-0003-2894-3900

https://doi.org/10.1016/j.csbj.2024.10.019
Received 17 August 2024; Received in revised form 7 October 2024; Accepted 13 October 2024
Available online 16 October 2024

proven to be quite ineffective when reaching the phase of refractory celiac disease [5]. Therefore, pharmacological treatment is essential for managing disease progression and helping patients obtain a healthier lifestyle, both physically and psychologically.

Studies on the human leukocyte antigen DQ2 (HLA-DQ2) gene led to the discovery of the role of tissue transglutaminases [6] in gluten-related diseases. The autoimmune response that leads to the symptoms experienced by celiac patients is primarily driven by Transglutaminase 2 (TG2), which catalyzes the cross-linking or deamidation of glutamine residues in gluten. Inhibition studies in search for novel drug-like molecules with high binding affinity to TG2 have been conducted using the traditional methods [7] of in vitro screening [8] through preclinical and clinical trials. Issues have arisen with targeting TG2 due to its important role in cell function processes [9]. Furthermore, the cost and time required for traditional laboratory techniques have compounded the challenges of conducting trials on promising molecules that inhibit the catalytic mechanism. It is noted that TG3 is also a novel target in dermatitis herpetiformis, an autoimmune skin condition associated with celiac disease [10].

Pharmaceutical companies have developed potential treatments by targeting inhibition with small molecules, gluten-breaking enzymes, octapeptides, and monoclonal antibodies [11]. However, many of these candidates were discontinued during clinical trials due to limited subject availability, insufficient protection, or lack of potential for optimized treatment [12]. Specific candidates designed for direct inhibition of TG2 have proven particularly promising in pre-clinical trials as they target the pathogenic process of the disease [13]. One TG2 inhibitor, GSK3915393, designed by GlaxoSmithKline [14], was discontinued in phase 1 of clinical trials after the decision to drop studies on celiac. Nevertheless, inhibition of the TG2 target has continued as the most widespread method for pharmacologists, highlighting the need to develop drugs against the protein.

Computational drug design has led the way in accelerating drug screening, utilizing various analytical frameworks to train machine learning algorithms. Machine learning has been widely used for quantitative structure-activity relationship (QSAR) [15] modeling, a computational approach that relates molecular structure and activity using supervised statistical methods. The bioactivity of a molecule can be measured by its ability to inhibit (categorical) or its inhibitory concentration (numerical). Ensemble learning and deep learning have been at the forefront of this shift, often being combined to create algorithms that leverage the strengths of both approaches. Certain weighted ensemble machine learning models like gradient and adaptive boosting classifiers can possibly prove promising when working with smaller bioassay datasets. These models can be implemented with weak learners, classifiers that generally perform slightly better than random guessing, to optimize speed while still achieving high accuracy [16,17].

In this study, we will use machine-learning approaches to facilitate the hit identification phase (Fig. 1A). Two machine-learning models will be developed to predict the bioactivity and potency of a new molecular candidate for inhibiting TG2 (Fig. 1B). Molecular fingerprints and descriptors will be extracted as features for the training data. The bioactivity model will be a binary classifier trained on a wide variety of active, inactive, and inconclusive molecules from lab bioassay data while an IC50 ranking model will be trained on molecules below a certain potency to obtain an algorithm that can be used to compare two candidates based on a ranking correlation coefficient. Feature importance analysis will be conducted to provide insight into the molecular features that either influence bioactivity positively or negatively. Researchers can prioritize the most promising candidates by effectively utilizing machine learning for activity prediction and potency estimation, leading to a faster and more efficient drug development process. Ultimately, these models will play a crucial role in designing more targeted and effective therapeutics during lead optimization and clinical trials, benefiting patients with gluten-related diseases.

2. Methods

2.1. Data preprocessing

Inhibitory data for transglutaminase 2 was obtained from the PubChem NCBI database's bioassay-derived bioactivity reports on *Protein-glutamine gamma-glutamyltransferase 2* [18] containing 1130 compounds (as of August 2023), including duplicates (See Supporting Information). In addition to identifiers and assay descriptions, each molecule contained fields pertaining to its performance, including potency values, a three-class uncertainty metric, and its classification. The ligands were listed as either active, inactive, unspecified, or ''inconclusive''. Around 55 % of molecules were active, while the majority of the remaining data was unspecified. Duplicates were removed based on the PubChem CID identifier, reducing the working dataset to 672 molecules. By slicing the active and inconclusive datasets based on the IC50 potency metric provided for these candidates, 336 molecules were selected for classification model development. Active molecules were sliced to only allow for molecules under a certain IC50 value while inconclusive molecules were sliced to accommodate those that had potencies closer to the inactive molecules in the original dataset. Each qualifying class was graphically analyzed via a histogram of the frequency of half-maximal inhibitory concentrations (IC50 [μM]) values within each subinterval.

2.2. Molecular descriptor & fingerprint calculation

Structure features were extracted from RDKit [19] fingerprint and descriptor modules to perform QSAR modeling. To obtain the SMILES

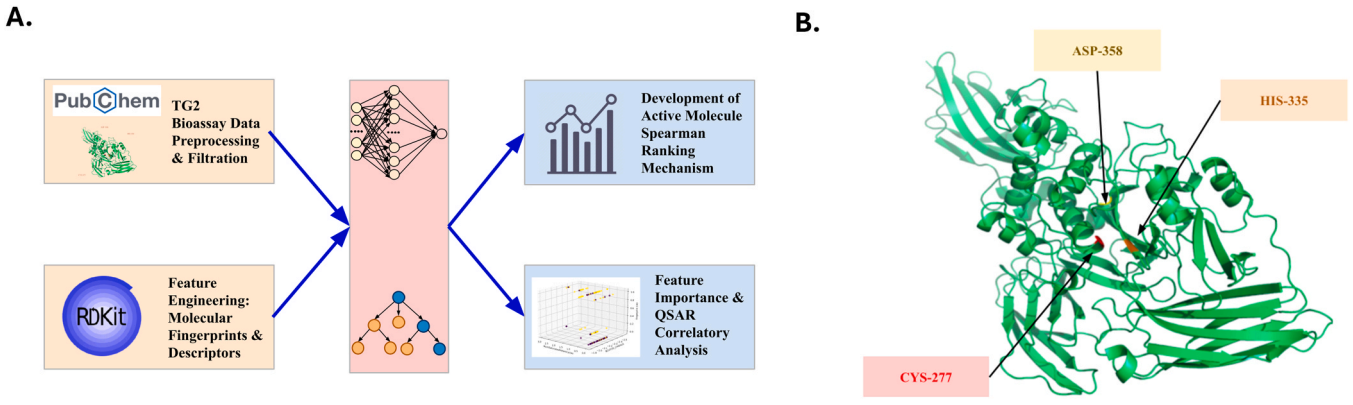

Fig. 1. A. Machine learned-based hit identification workflow pipeline. Data preprocessing, model training, and post-evaluation procedures **B.** Transglutaminase 2 structure and its catalytic triad responsible for the protein's significant process and facilitation.

representations of all molecules in the dataset, the PubChem Rest API (PubChempy) was utilized. Four molecular fingerprint libraries provided by RDKit were tested (Avalon, MACCS Keys, Morgan, RDKit) [20] on various general binary classifiers and sequential artificial neural network models. The RDKit Chem, AllChem Avalon, and rdMolDesriptors modules provided the functions necessary for swiftly calculating these fingerprints.The use of Avalon and RDKit fingerprints had the best testing set performances but RDKit distinguished itself in providing selective representations of the substructures associated with each fingerprint bit vector allowing for further investigation of feature importance. 457 fingerprints and 210 descriptors were calculated and added to the training set. Feature elimination on the generated fingerprints was conducted by elimination of fingerprints with variance threshold ≤ 0.1 (Eq. 1) and groups of fingerprints with Pearson correlation coefficient ≥ 0.9 (Eq.2) before the final dataset was normalized with a standard scaler (Eq. 3).

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1}$$

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}} \tag{2}$$

$$X_i' = \frac{X_i - \mu}{X_{std}} \tag{3}$$

$\overline{x}$: average value of x

x_i: ith value of x

X_{std}: standard deviation of X

2.3. Tanimoto similarity analysis

To evaluate the training potential of the biological assay and the complex diversity of molecules in the dataset, a comprehensive similarity analysis was performed on all molecule cross-relationships using the fingerprint bit vector-based Tanimoto chemical similarity coefficient [21] (Eq. 4), a widely used metric in the field of cheminformatics. The coefficient determines how similar two molecules A and B are based on their molecular fingerprint bit vectors. For instance, a Tanimoto coefficient of 0.85 indicates that the molecules are 85 % similar in structure. RDKit's DataStructs module was used to acquire tools for similarity calculations. Typically, the number of clusters of highly similar molecules should be limited for training to ensure a potential model can generalize on a large variety of inputs. The results were visualized, and it was determined that the dataset was sufficiently chemically diverse to train an efficient supervised classification model.

$$T_{A,B} = \frac{\sum_{j=1}^{n} x_{jA} x_{jB}}{\sum_{j=1}^{n}(x_{jA})^2 + \sum_{j=1}^{n}(x_{jB})^2 - \sum_{j=1}^{n} x_{jA} x_{jB}} \tag{4}$$

2.4. QSAR ensemble model architecture

Binary classifiers from scikit-learn 1.2.2 encapsulated into the lazy predict classifier and also external boosting algorithms were evaluated on the initial testing set to obtain a highly tunable model that can be optimized with hyperparameter tuning. Some of the boosting packages, like CatBoost 1.2.2 & XGBoost 2.0.3, came from external libraries, but the highest-performing boosting model, AdaBoost (adaptive boosting), was chosen for hyperparameter optimization. After testing numerous weak learners for the adaptive boosting algorithm, including tree-based methods and simple deep learning models, a multi-layer-perceptron deep learning adaptive boosting ensemble and a low-depth random forest ensemble were built, tuned, and finalized as deployable models. For comparison, Chemprop's message-passing graph neural networks were trained and tuned as well. 5-fold cross-validations were performed over 50 trials for each type of model. The imbalanced dataset required that a stratified split be performed. Classification performance metrics [22], including Precision, Recall, F1, Accuracy, MCC, and AUC-ROC (Eqs. 5–9), were recorded and visualized.

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{5}$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{6}$$

$$F1 = \frac{2(Precision)(Recall)}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{8}$$

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{9}$$

T_P : True Positives

T_N : True Negatives

F_P : False Positives

F_N : False Negatives

2.5. Graphical feature importance analysis

A random forest was utilized to determine important fingerprints and descriptors responsible for predicted bioactivity. Pair plots and correlation matrices, along with 3D plots, were used to evaluate QSAR elements. The substructures represented by the top fingerprint bit vectors and descriptors with the highest feature importance were analyzed for structure-descriptor-activity correlations. Each fingerprint's respective substructure was analyzed to establish molecular properties necessary for TG2 inhibition.

2.6. Spearman coefficient IC50 ranking study

To allow for comparison between two potential candidates for further study, a model for the prediction of IC50 of molecules with prioritization on the correct ranking of candidates. The Spearman coefficient ranking coefficient (Eq. 10) was chosen to evaluate the performance of testing set predictions. A new training set was constructed from the original unbalanced data set using molecules with potency $\leq$ 10 μM to generalize the model on primarily active molecules. A 0.2 variance threshold, low-depth random forest feature importance, and recursive feature elimination were used to cut features from the dataset to 200. Models that exist in the regressor package of lazypredict (lazyregressor) were utilized to train scikit-learn models and external boosting algorithms were trained. The metrics recorded were the Spearman rank correlation coefficient (Eq. 10), root mean squared error (Eq. 11), mean absolute error (Eq. 12), and the coefficient of determination (R-squared) (Eq. 13).

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} \tag{10}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i' - y_i)^2}{n}} \tag{11}$$

$$MAE = \sum_{i=1}^{n} \frac{|y_i' - y_i|}{n} \tag{12}$$

$$R^2 = 1 - \frac{\Sigma(y_i' - y_i)^2}{\Sigma(y_i - \bar{y})^2} \quad (13)$$

2.7. Web app deployment

Using a Flask backend and HTML/CSS/JS frontend, a web app (Celiac Informatics) was constructed to aid the classification of potential TG2 inhibitors, and the features associated with novel potency. The app was built to take the input of a molecule in its SMILES format and output the following four analytics: the predicted bioactivity of the molecule against TG2, a summary of its important fingerprints of descriptors, a short evaluation of its drug-likeness with five of the most conventionally used rules, and the relative IC50 ranking for active molecules. For the drug-likeness score, Lipinski's rule of 5 states that oral drugs should have less than or equal to 5 hydrogen bond donors [23], a molecular weight less than or equal to 500 Da, a logP less than or equal to 5, and less than or equal to 10 hydrogen bond donors. Ghose requires a molecular weight and logP in the inclusive domains of [160,480] Da and [−0.4, 5.6] Da, respectively. It also allows for a relative molecular mass in [40,130] inclusively and no more than 70 atoms and no less than 20 atoms in the molecule [24]. Egan's rule is more flexible with concrete structural properties and only requires a logP less than or equal to 5.88 and a topological polar surface area less than or equal to 131.6 [25]. Muegge's rule [26] is the most conservative, with requirements in nine categories. Its molecular weight and logP ranges are [200,600] and [−2, 5], respectively. In addition to requiring a topological surface area less than or equal to 150 and no more than 7 rings, the molecule must have at least 5 carbon atoms and 2 heteroatoms as well no more than 15 rotatable bonds. It should also have less than or equal to 10 hydrogen bond acceptors and 5 hydrogen bond donors. Heroku was used to deploy a web service for the application: https://celiac-informatics-v1–2b0a85e75868.herokuapp.com.

3. Results & discussion

3.1. IC50-based data slicing & Tanimoto similarity evaluation

The goal of this exploratory data analysis was to obtain a moderately balanced training set by evaluating the potencies of molecules by the four classes: active, inactive, unspecified, and inconclusive. The inconclusive class was deemed unfit for this analysis due to the lack of recorded potency values and error metrics in their respective lab assays. Out of 424 potential active inhibitors and 177 unspecified inactive candidates, we performed slicing based on their IC50 values. Molecules with an IC50 $\leq$ 1 μM are typically the most potent and suitable for further drug development [27]. The active dataset was created by applying an IC50 threshold of less than 0.6 μM, based on their distribution, as illustrated in Fig. 2A. This refinement resulted in approximately 50 % of all active molecules being selected. As a result, 216 active molecules were selected, along with 75 unspecified molecules (see Fig. 2B). These unspecified molecules were combined with 45 documented inactive molecules to create a training set comprising 120 inactive molecules. This allowed for a desirable ratio between the active and inactive classes, providing a baseline performance of approximately 0.64 accuracy and 0.50 AUC for model training. To ensure the potential model can be trained to make generalized predictions on various samples from molecular databases, the molecular similarity was analyzed using the Tanimoto coefficient (Fig. 2C). The generated molecular fingerprint bit-vector strings of each molecule were quantitatively compared to generate a heatmap to visualize the prominence of high similarity clusters. The dataset was determined to be sufficiently diverse and could be split using stratification.

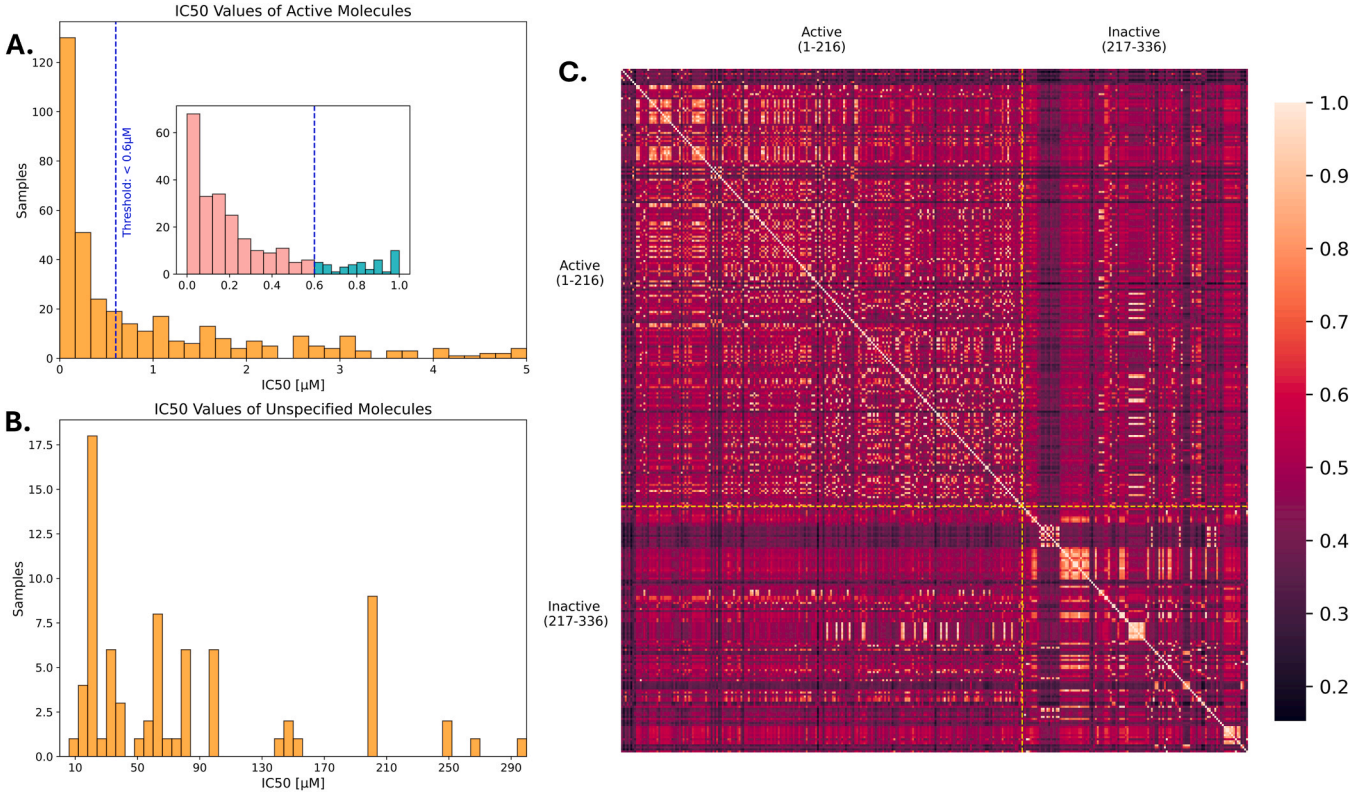

Fig. 2. Histograms of IC50 values of the candidate molecules against TG2. **A.** The active candidates graphed in the main figure range from 0 to 5 μM. The inset illustrates a histogram of molecule potency < 1.0 μM. Most active molecules lie in the range displayed by the inset, and within this domain, 0.6 μM is a suitable place to slice. **B.** Unspecified candidates range from 10 to 1000 μM (only those < 300 μM are shown in the figure; the unspecified set was sliced at this point). **C.** Tanimoto coefficient molecular similarity distribution.

3.2. The performance of MLP & low-depth random forest boosting ensembles

The top-performing models using the default hyperparameters on an initial testing set are shown in Table 1 (the full list can be found in Table S1). The training-to-testing set ratio was 75:25. Boosting models like AdaBoost and LightGBM outperformed tree classifiers, support vector machines, and other sophisticated methods for the initial task of finding an algorithm for hyperparameter tuning. AdaBoost was implemented with the conventional stump-based decision tree; it performed with an average accuracy of 96 % on the initial testing set, as a result of its effective use of a chain of weak learners and weight resampling. The use of weak learners in ensemble learning methods and its self-correcting philosophy were instrumental to its performance (Fig. 3) [28,29].

Bioactive molecules possess synergistic features that relate to their binding affinity and interactions with target proteins. The base learner in a boosting ensemble is designed to use a chain of weak estimators, hence the term "weak learner." A stump is a version of a weak learner that predicts outcomes based on a single feature. However, a stump estimator can be extended into a small, low-depth tree or even a forest of such trees, known as a low-depth random forest. This method can efficiently predict bioactivity by leveraging numerous fingerprint-descriptor relationships, making it more effective than using a simple stump alone. Additionally, a multi-layer perceptron can serve as a deep learning weak learner if it is constrained to a few hidden layers, thus preserving its intended weakness and maintaining quick estimator performance. A base learner algorithm can be integrated as the base estimator into the adaptive boosting pipeline (Fig. 3).

Two models were set up, one with a random forest base learner, and another with a multi-layer-perceptron base learner. Their hyperparameters were tuned using Bayesian optimization based-package Optuna [31,32] on a few of the parameters of the base learner as well as some general parameters from the adaptive boosting algorithm itself. For example, the random forest max depth and minimum sample split were the tuned hyperparameters. For the multi-layer perceptron with one hidden layer, the hyperparameters tuned were: the number of hidden layer neurons, the activation function, the initial learning rate, and the maximum number of iterations. The adaptive boosting-specific parameters that were tuned in the methods were the number of estimators and the learning rate. The parameters for each model that optimized the accuracy were selected for final evaluation.

The final adaptive boosting models were constructed and evaluated using stratified 5-fold cross-validation to determine whether they could maintain their performance when trained on 5 different subsets of the original cleaned dataset. A random forest [33] with a maximum depth of 3 and 3 minimum sample splits was constructed as the base estimator in one adaptive boosting method which had 442 estimators and a learning rate of about 0.996. It performed at an average of ~92 % accuracy during the sample 5-fold cross-validation (Table 2). Furthermore, a multi-layer-perceptron (MLP) [34] was also implemented as a weak learner for the second adaptive boosting method and it had the following parameters: one hidden layer containing 50 neurons, a relu activation function, an initial learning rate of 0.001, and 1000 maximum iterations. The adaptive boosting pipeline had 279 estimators and a learning rate of about 0.0230. Its backpropagation system, along with adaptive boosting weight resampling, allowed for the increase in performance on initially incorrectly predicted samples; its cross-validation average accuracy was ~93 % (Table 2). For comparison, Chemprop's message-passing graph neural networks [35,36] were also cross-validated and achieved an accuracy of ~89 %, exhibiting long testing times and difficulty within its trials (See Table S2 for full results). The multi-layer perceptron adaptive boosting ensemble was selected as the final model for the web app.

Table 1
Top performing bioactivity classification algorithm on the initial testing set.

Model	Accuracy	AUC	F1 Score	Precision	MCC
Stump-Based Adaptive Boosting Ensemble (AdaBoost)	0.96	0.96	0.95	0.96	0.92

3.3. Individual molecular fingerprint and descriptor importance analysis yields structural significance

Through feature importance analysis from a random forest estimator trained on a comprehensive dataset of 336 molecules, we identified essential molecular descriptors & fingerprints for bioactivity. These descriptors encompassed the count of saturated heterocycles, saturated rings, and properties associated with Van der Waals surface area contributions, surface molecular recognition, orbital electronegativity, and charge distribution. Due to the distinct nature and dependencies of each descriptor, it was essential to identify specific substructures highlighted by RDKit fingerprints that were prevalent in active molecules. Most essential fingerprints were in the RDKit fingerprint domain of 300–500 (See Fig. 4A & 4C); 6 were identified by the feature importance algorithm as potential drivers in activity classification. The frequency of specific substructures, represented by the nine fingerprints with the highest importance, was analyzed across all molecules in the dataset. This analysis helped identify whether each fingerprint was more prominent in active or inactive molecules, thereby determining their respective associations. Furthermore, the most important descriptors are illustrated in Fig. 4D, while the correlation between these descriptors and bioactivity is summarized in Fig. 4B.

The fingerprint with the highest feature importance, fingerprint 495 (see Fig. 4C), is the Michael acceptor group (see Fig. 5A). Michael acceptor groups are key components in many covalent inhibitor drugs [37]. They facilitate the Michael addition reaction, which is a key chemical process that occurs when the molecule binds to its target. This reaction involves the addition of a nucleophile to an α,β-unsaturated carbonyl compound, resulting in the formation of a new covalent bond. The presence of a Michael acceptor group in a terminal position enhances its ability to form strong interactions with biological targets via a Michael addition reaction [38], thereby increasing its binding affinity and bioactivity. Therefore, fingerprint 495 could be considered an essential component of a molecule that could inhibit TG2. There are 180 active molecules that contain fingerprint 495 while only 31 inactive molecules have it (Fig. 4A). We found that the position (terminal or non-terminal) of the Michael acceptor group could possibly explain the difference between some active and inactive molecules. For example, Fig. 5A illustrates an active molecule featuring a Michael acceptor group in a terminal position. In contrast, Fig. 5B shows an inactive molecule where the Michael acceptor group is in a non-terminal position. Both molecules exhibit fingerprint 495; however, the inactive molecule demonstrates insufficient binding affinity due to the non-terminal position of the Michael acceptor group. Notably, 14 out of the 31 inactive molecules with fingerprint 495 either have a slightly different substructure or feature a Michael acceptor in a non-terminal position, which may affect their binding efficacy.

3.4. Multi-feature correlation and synergism analysis

The fingerprint and descriptor features could have synergistic effects on bioavailability. The first important relationship is between the number of saturated rings and heterocycles (Fig. 6A). Most active molecules had one saturated heterocycle and one saturated ring, and most of them had the same substructure. Adding more rings does not necessarily decrease the molecule's bioactivity as shown in the plot. In addition, the data supports that adding more saturated rings when maintaining a single heterocycle can maintain bioactivity. However, drug-likeness must also be considered when designing a molecule, for instance, the

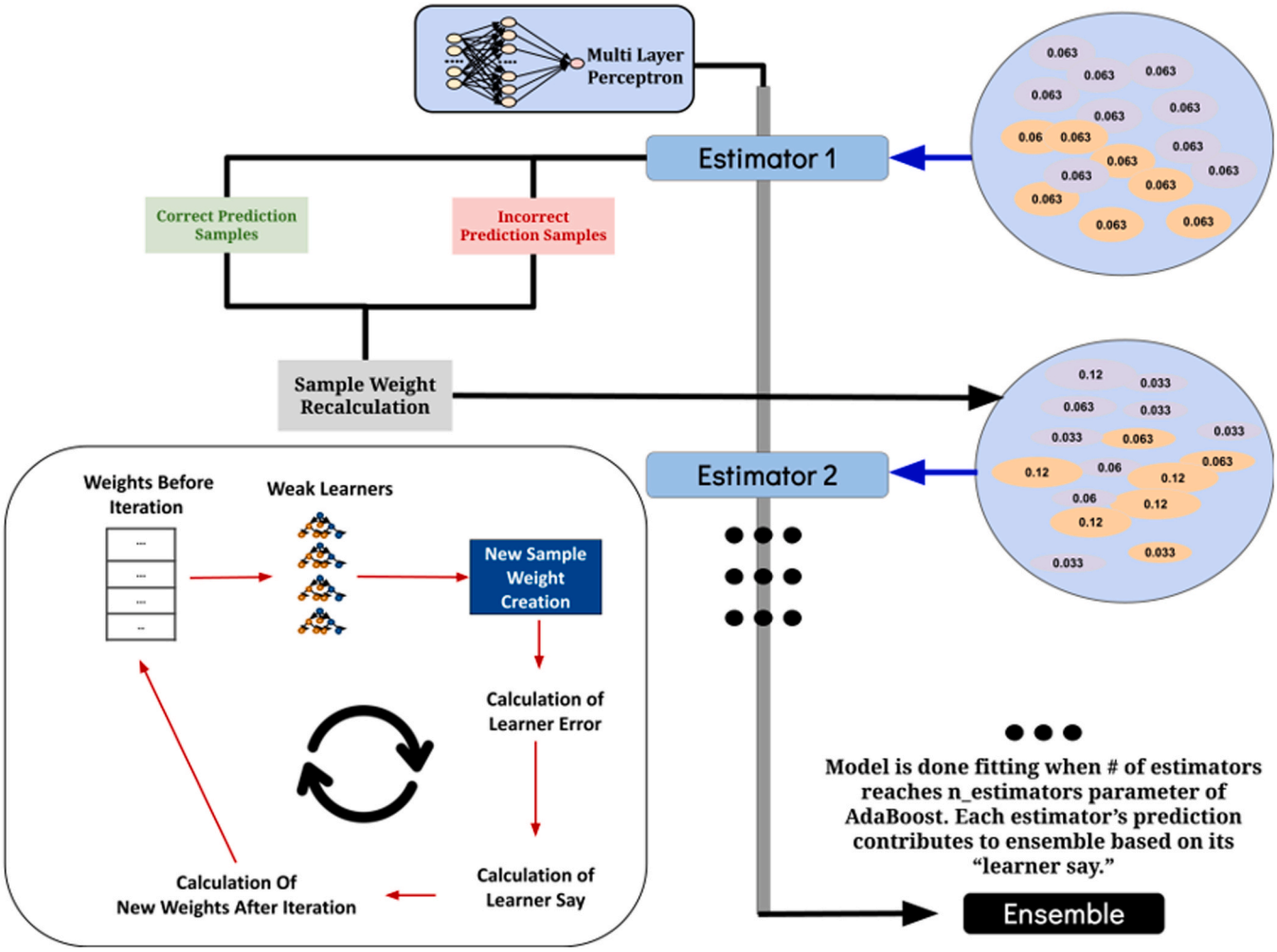

Fig. 3. The architecture of the final adaptive boosting ensemble algorithm with a multi-layer perceptron weak learner [30]. This weak learner is the backbone of the iterative algorithm, which is trained and tested successively throughout the run. Based on the performance of a particular estimator in the chain, new weights are assigned to improve predictions on incorrect prediction samples. The learner error and the new weights must be calculated on each iteration. The final tuned model contains a chain of estimators, and a weighted average of the prediction is calculated to determine the final classification.

Table 2
Final Optimized Model Performance (5-Fold Cross Validation over 50 Trials).

Model	Accuracy	AUC	F1 Score
Multi-Layer Perceptron Adaptive Boosting Ensemble	0.93 ± 0.02	0.91 ± 0.02	0.95 ± 0.02
DummyClassifier (Baseline)	0.64	0.50	0.50

maximum seven-ring requirement of Muegge's rule. The structure-activity relationship can be further investigated using a pair plot. From Fig. 6B, there is a certain range (50−250) for the descriptor SlogP_VSA2 that exhibits a higher chance for a molecule to be bioactive when the number of heterocycles is ≥ 1.

Some molecular fingerprints tokenize varieties of substructures that influence the values of certain count-based properties like the number of recorded saturated rings and heterocycles, and also chemical property descriptors [39]. BCUT2D partial charge descriptors highly influence the potential for stronger intermolecular forces between the molecule and the binding site [40]. By observing pair plots of these relationships, generally, a molecule is more likely to be active if it has more heterocycles and a lower BCUT2D_CHGLO or higher BCUT2D_CHGHI value (Fig. 6C-D). BCUT2D_CHGLO and BCUT2D_CHGHI have a negative correlation (−0.52) (Fig. 4B).

QSAR analysis can be further enhanced by searching for fingerprints that have some relationship with structural and/or chemical property descriptors. From Fig. 7A-B, it can be observed that the presence of pivotal molecular fingerprints, 193 and 263, combined with more negative BCUT2D_CHGLO or more positive BCUT2D_CHGHI and the number of saturated heterocycles yields more active molecules. This may result from the fact that many of the substructures represented by these fingerprints contain a heteroatom that is part of a saturated heterocycle. For example, piperazine rings, which are commonly present as substructures in small pharmaceutical compounds (Fig. 8A), contain two nitrogen atoms in a six-membered cyclic structure. The ring contains solely single bonds giving it the "saturated" characterization, but it also has strong pharmacokinetics properties specifically in the fields of oral bioavailability and solubility. Particularly, when a nitrogen atom in these rings is connected to another non-carbon atom like sulfur, it can have more substantial medicinal effects [41]. The substructure represented by fingerprint 193 and 263 in Fig. 8A is a sulfenamide group and the active molecule contains the oxidized form (sulfonamide) [42]. This can explain why fingerprints like 193 or 263 represent some substructures that are not necessarily heterocycles but merely contain a heteroatom bonded to a substantial element or functional group. A single molecular fingerprint bit does not need to represent a unique substructure for all molecules, but it can constitute a broad structural property present in all molecules. Furthermore, when combined with another important structural descriptor like the number of saturated

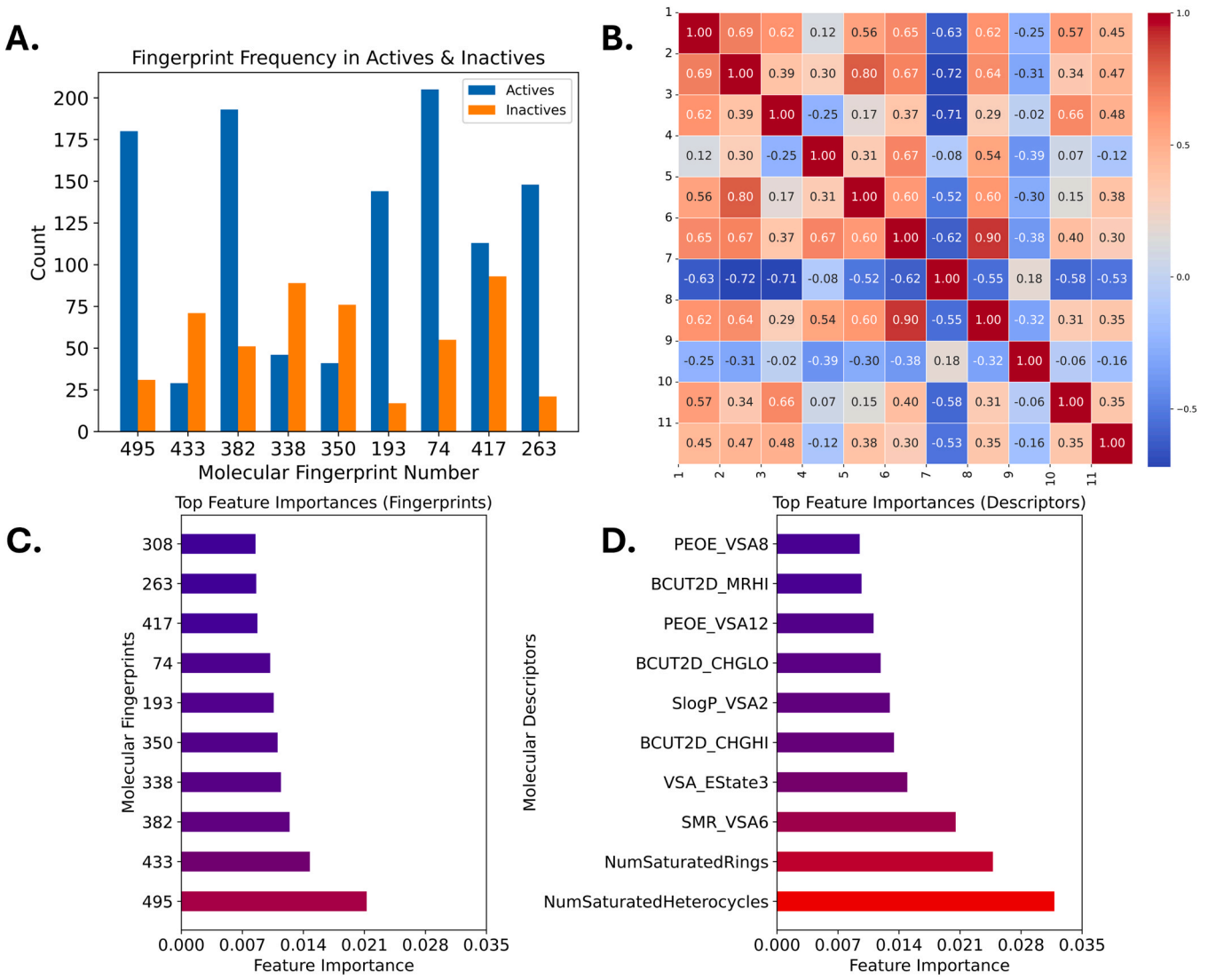

Fig. 4. A. Frequency distribution of the nine highest feature importance RDKit molecular fingerprints in classified molecules of the dataset. The fingerprints represent certain substructures of a molecule that could be indicative of properties that support bioactivity **B.** Correlation matrix of ten highest feature importance molecular descriptors (1. NumSaturatedRings, 2. NumSaturatedHeterocycles, 3. SMR_VSA6, 4. VSA_Estate, 5. BCUT2D_CHGHI, 6. SlogP_VSA2, 7. BCUT2D_CHGLO, 8. PEOE_VSA12, 9. BCUT2D_MRHI, 10. PEOE_VSA8) **C-D.** Visualized feature importance rankings for fingerprints and descriptors.

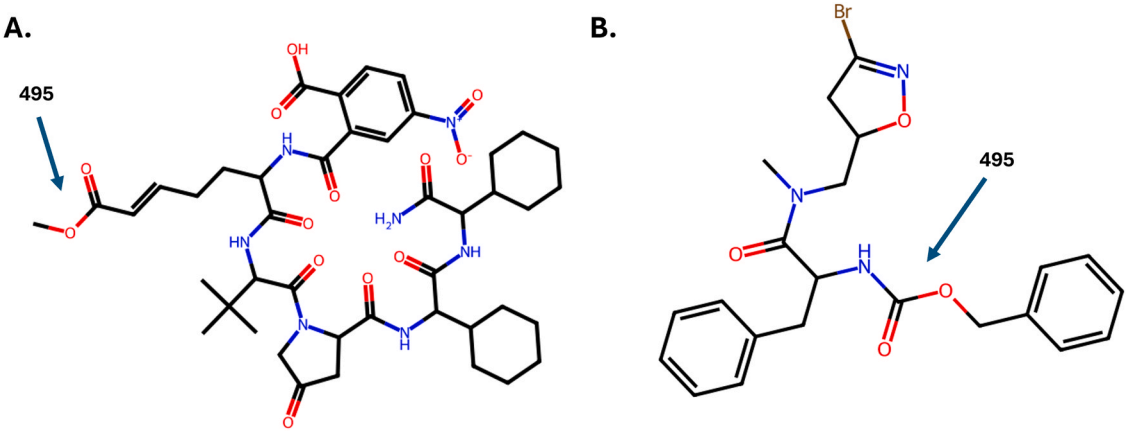

Fig. 5. A. An active molecule (2-[[(E,2S)−1-[[(2S)−1-[(2S)−2-[[(1S)−2-[[(1S)−2-amino-1-cyclohexyl-2-oxoethyl]amino]−1-cyclohexyl-2-oxoethyl]carbamoyl]−4-oxopyrrolidin-1-yl]−3,3-dimethyl-1-oxobutan-2-yl]amino]−7-methoxy-1,7-dioxohept-5-en-2-yl]carbamoyl]−4-nitrobenzoic acid) with a Michael acceptor (495) in a terminal position. **B.** An inactive molecule (Dihydroisoxazole, 2b) with a Michael acceptor (495) in a non-terminal position.

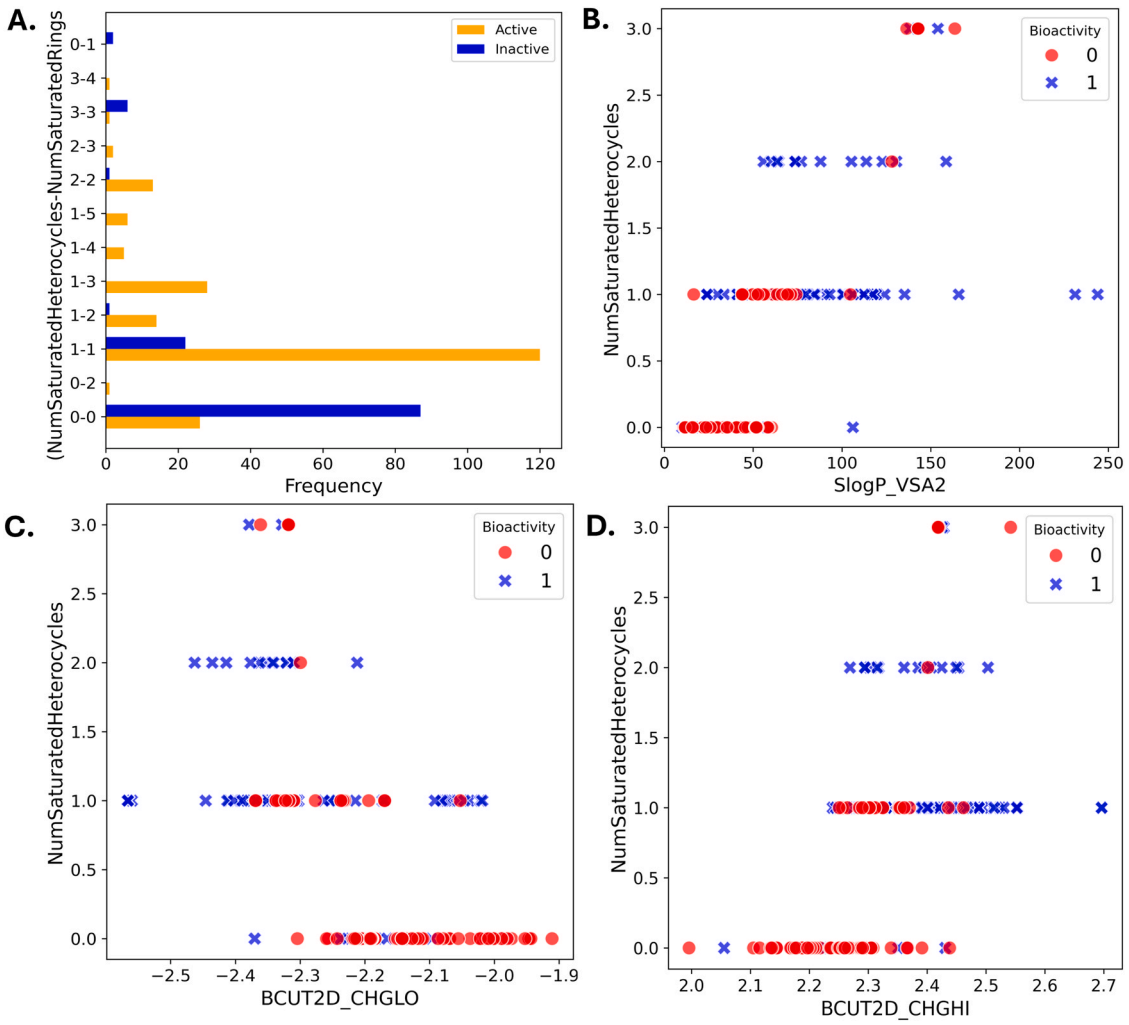

Fig. 6. **A.** The frequency of the number of saturated heterocycles and rings in active & inactive molecules. **B.** Activity pair plots of the number of saturated heterocycles and SlogP_VSA2. **C-D.** Activity pair plots of the number of saturated heterocycles and two BCUT2D descriptors.

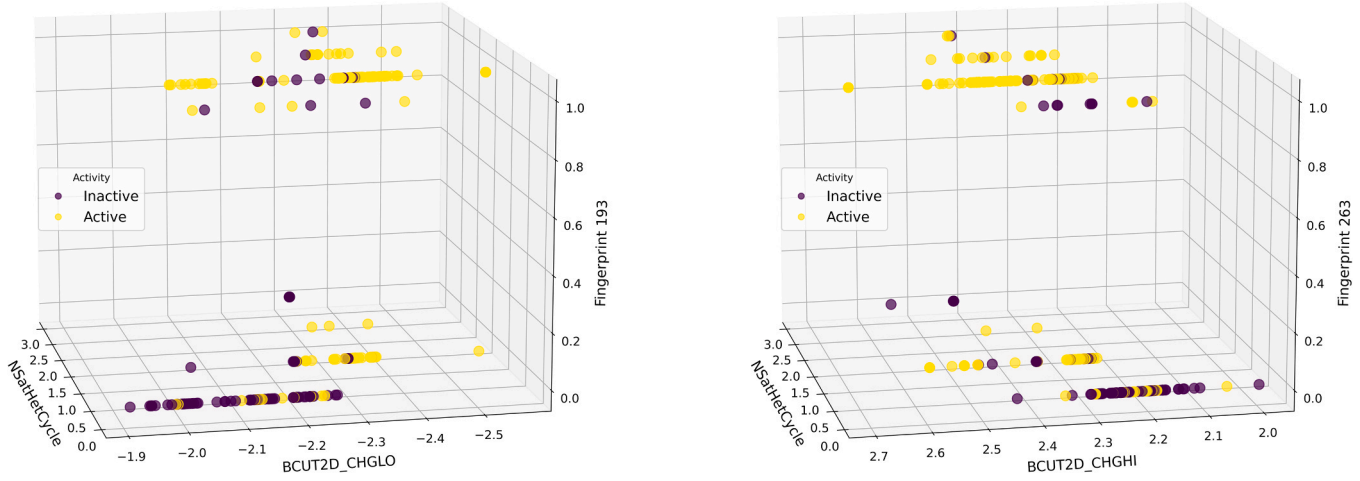

Fig. 7. Each axis represents one of the features in the model: a molecular fingerprint, which encodes the presence or absence of particular substructures; a structural descriptor, which captures geometric and topological properties; a property descriptor, which reflects physicochemical properties. By plotting molecules in this 3D space, clustering patterns can be observed to distinguish active molecules from inactive ones. This visualization aids in identifying key features associated with molecular activity, highlighting regions in the descriptor space where active compounds are concentrated **A-B.** Activity triple-plots of structural RDKit fingerprints, saturated heterocycle count, and BCUT2D descriptors.

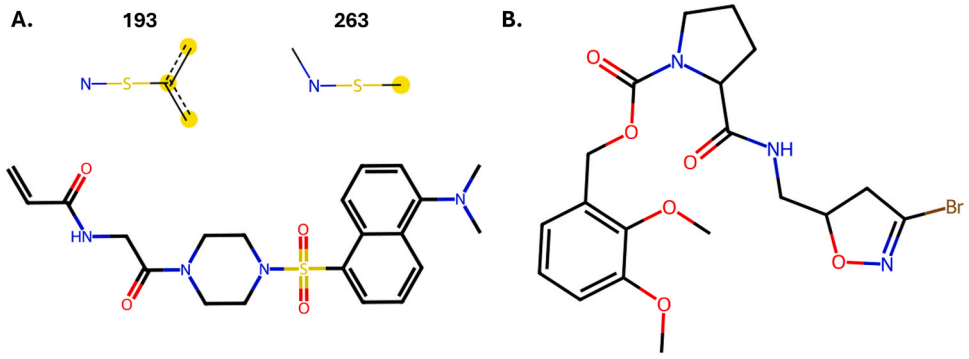

Fig. 8. A. An active molecule (N-[2-[4-[[5-(dimethylamino)−1-naphthyl]sulfonyl]piperazin-1-yl]−2-oxo-ethyl]prop-2-enamide (3 h)) with a heteroatom connected to a sulfur atom, supporting its medicinal effect **B.** An inactive molecule ((2,3-dimethoxyphenyl)methyl (2S)−2-[[(5S)−3-bromo-4,5-dihydro-1,2-oxazol-5-yl] methylcarbamoyl]pyrrolidine-1-carboxylate) with a heteroatom connected to a carbon.

heterocycles, it can produce synergistic effect. Furthermore, it can explain inactive points within the plot that happen to have a heterocycle. The heteroatoms within the heterocycle(s) of such molecules may only be connected to simply another carbon atom and hence do not necessarily have the properties indicative of biological activity (Fig. 8B). Inactive molecules with a heterocycle that do not fall under this characterization may contain other driving substructures and properties, and this would warrant the need for a more comprehensive investigation of other important fingerprints.

3.5. Spearman coefficient Performance of predictive IC50 models

The boosting ensemble classification model is highly effective for rapidly screening potential drug-like molecules. However, for comparing candidates based on IC50 values, regression methods are more suitable. Predicting IC50 values accurately can be challenging due to the small-scale dataset. Nonetheless, mathematical coefficients like the Spearman coefficient can help evaluate a model's ranking capability.

Boosting ensemble regressors can be employed to develop a model with a high Spearman coefficient. Traditional supervised models were also evaluated, with gradient boosting ensembles and support vector machines demonstrating the best performance in optimizing the Spearman coefficient (see Table 3 for top models and Table S3 for the full list).

CatBoost and support vector machine had spearman coefficients of 0.72 indicating solid preservation of ranking orders. At the same time, their predictions need not be taken literally as seen from their poor accuracy metrics but should be primarily used for the comparison of two or more molecules. In essence, if the IC50 predictions for two molecules from these models are being compared, the molecule with the lower prediction will likely be more potent and effective but their actual values are not necessarily quantitatively related. CatBoost performs better in terms of accuracy for molecules with low actual IC50 values, but its performance declines as IC50 values increase (see Fig. 9A). On the other hand, Support Vector Machine (SVM) predictions are generally lower than the actual IC50 values, and this discrepancy becomes more pronounced with higher IC50 values. This consistent underestimation by SVM contributes to its poorer performance metrics, including higher RMSE, MAE, and lower R-squared values, as shown in Fig. 9B.

Table 3
Performance of IC50 potency regression algorithms.

Model	Spearman Coefficient	Root Mean Squared Error	Mean Absolute Error	R-Squared
Ordered Gradient Boosting Ensemble (CatBoost)	0.72	2.36	1.49	0.31
Support Vector Machine	0.72	2.73	1.54	0.08

3.6. Web application: celiac informatics

The web application predicts bioactivity classification for novel molecules against TG2 using the multi-layer perceptron adaptive boosting ensemble model as its backend and outputs important descriptors or fingerprints using RDKit's bit visualizations. On taking the SMILES notation of a potential ligand, the algorithm is run within a few seconds before redirecting to a page with a bioactivity report. To ensure that the input molecules exhibit the qualities of a potential drug, it checks the classification of the model as organic and sets a required molecular weight of at least 60 Da. Additionally, a molecule's bioactivity is not predicted unless it is considered sufficiently drug-like as per the 5 drug-likeness rules (Lipinski, Egan, Muegge, Veber, and Ghose). Repeated violations of the pharmacokinetic properties detailed within these rules may cast the molecule as "undrug-like" and not worthy of bioactivity prediction. The molecule's relative IC50 positional ranking is also graphed in the scatterplot along with all the active molecules in the original dataset to allow for comparison and evaluation of potency.

4. Conclusion

This study employed a combination of computational QSAR and ensemble learning techniques to develop a high-accuracy model for predicting drug bioactivity and analyzing key features of active molecules targeting TG2 in celiac disease. Various methods, including boosting, support vector machines, tree-based algorithms, and basic deep learning techniques, were tested to create a robust model for molecular activity prediction. Boosting ensembles, such as adaptive boosting and gradient boosting, demonstrated the highest performance in activity classification tasks. The original adaptive boosting model was refined by tuning the weak learner parameters, resulting in two efficient algorithms: one using a multi-layer perceptron and the other a low-depth random forest as base learners. Feature importance analysis highlighted the role of terminal Michael acceptor groups and sulfonamide groups as critical components of TG2-bioactive molecules, showing the high feature importance. A regression model, trained to maximize the Spearman coefficient, was also developed to rank potential inhibitors when precise potency measurements were unavailable. The cheminformatics web tool, Celiac Informatics (https://celiac-informatics-v1–2b0a85e75868.herokuapp.com) encapsulates these machine-learning algorithms, offering a user-friendly interface for exploring statistical data and understanding the descriptors that influence a molecule's performance in inhibiting TG2. The deployment of a

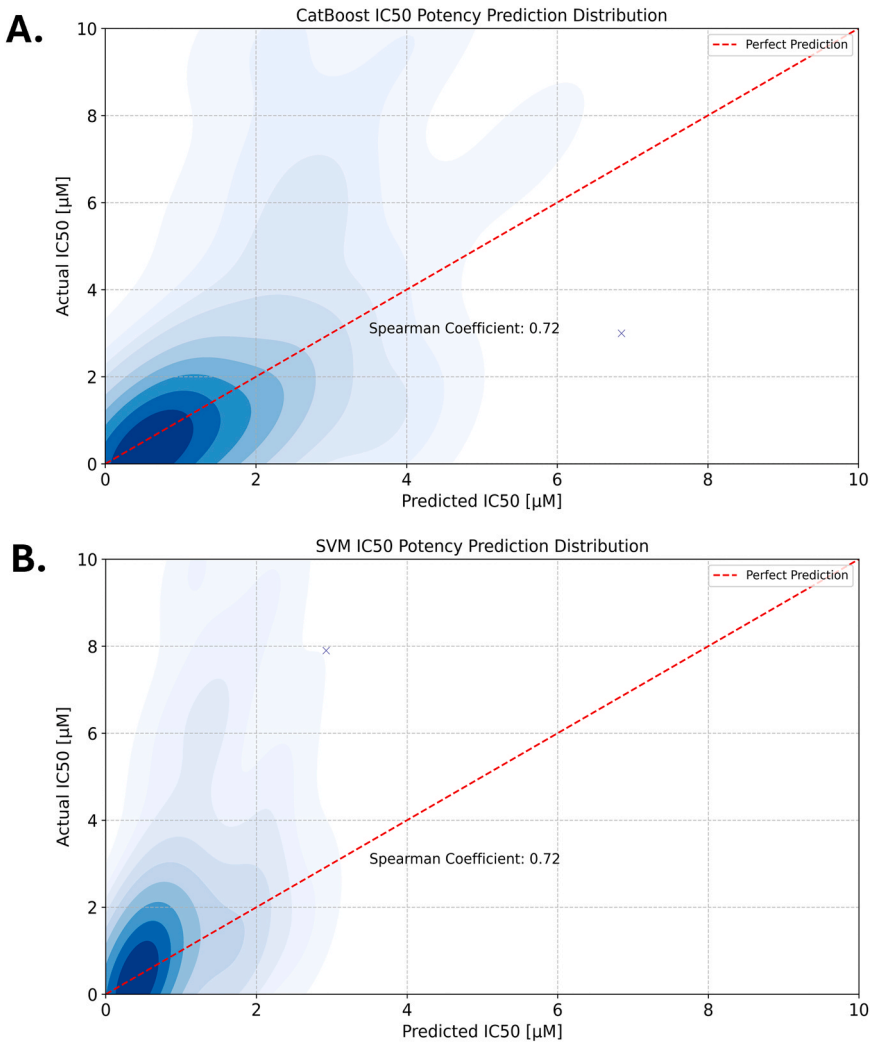

Fig. 9. A-B. IC50 Prediction Distribution of CatBoost & Support Vector Machine Models.

highly accurate boosting algorithm in a fast web application will assist medicinal chemists in swiftly designing potential derivatives against TG2 with drug-like properties.

CRediT authorship contribution statement

Ibrahim Wichka: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pin-Kuang Lai:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

There is no conflict of interest.

Acknowledgements

This work was financially supported by a start-up fund from Stevens Institute of Technology.

Supporting Information

This material is available free of charge via the Internet at https://www.sciencedirect.com/journal/computational-and-structural-biotechnology-journal.

- Table S1. Full list of bioactivity model classification results from lazypredict on the initial testing dataset with following metrics: Accuracy, AUC, F1 Score, Precision, MCC, Speed.
- Table S2. Full list of performance metrics for final MLP AdaBoost ensemble model with comparison to other sophisticated algorithms.
- Table S3. Full list of IC50 model regression results from lazypredict on the initial testing dataset with following metrics: Spearman rank correlation coefficient, Root mean squared Error, Mean Absolute Error, R-squared.
- TG2_compound_smiles.xlsx: All 1130 tested compounds for TG2 bioassays.
- TG2_classfication_train_test: All 336 filtered compounds for classification modeling.
- TG2_regression_train_test: All 424 compounds with IC50 $\leq$ 10 μM for regression modeling.
- Fingerprints_structures_final.xlsx: All 336 molecular structures from the original classification dataset and their respective important fingerprint substructure representations.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.019.

Data Availability

All public data sets (TG2_compound_smiles.xlsx), filtered data sets (TG2_classfication_train_test and TG2_classfication_train_test), and feature sets (Fingerprints_structures_final.xlsx) generated in this work are provided as Supporting Information. The Python notebooks for the classification and regression models and codes for the web app and the final saved models are available at https://github.com/Lailabcode/celiac-informatics. The Chemprop training code is available at https://github.com/ibrahimwichka/Chemprop_model_testing. The Celiac Informatics web app is available at https://celiac-informatics-v1–2b0a85e75868.herokuapp.com.

References

[1] Mahadov S, Green PHR. Celiac disease: a challenge for all physicians. Gastroenterol Hepatol 2011;7(8):554–6.
[2] Losowsky MS. A history of coeliac disease. Dig Dis 2008;26(2):112–20. https://doi.org/10.1159/000116768.
[3] van Berge-Henegouwen GP, Mulder CJ. Pioneer in the gluten free diet: Willem-Karel Dicke 1905-1962, over 50 years of gluten free diet. Gut 1993;34(11):1473–5. https://doi.org/10.1136/gut.34.11.1473.
[4] Makharia GK. Current and emerging therapy for celiac disease. Front Med 2014;1:6. https://doi.org/10.3389/fmed.2014.00006.
[5] Machado MV. New developments in celiac disease treatment. Int J Mol Sci 2023;24(2). https://doi.org/10.3390/ijms24020945.
[6] Paolella G, Sposito S, Romanelli AM, Caputo I. Type 2 Transglutaminase in coeliac disease: a key player in pathogenesis, diagnosis and therapy. Int J Mol Sci 2022;23(14). https://doi.org/10.3390/ijms23147513.
[7] Van Buiten CB, Elias RJ. Gliadin sequestration as a novel therapy for celiac disease: a prospective application for polyphenols. Int J Mol Sci 2021;22(2). https://doi.org/10.3390/ijms22020595.
[8] Chrobok NL, Bol JGJM, Jongenelen CA, Brevé JJP, El Alaoui S, Wilhelmus MMM, et al. Characterization of Transglutaminase 2 activity inhibitors in monocytes in vitro and their effect in a mouse model for multiple sclerosis. PloS One 2018;13(4):e0196433. https://doi.org/10.1371/journal.pone.0196433.
[9] Tempest R, Guarnerio S, Maani R, Cooper J, Peake N. The biological and biomechanical role of Transglutaminase-2 in the tumour microenvironment. Cancers 2021;13(11). https://doi.org/10.3390/cancers13112788.
[10] Kaunisto H, Salmi T, Lindfors K, Kemppainen E. Antibody responses to Transglutaminase 3 in dermatitis herpetiformis: lessons from celiac disease. Int J Mol Sci 2022;23(6). https://doi.org/10.3390/ijms23062910.
[11] Mitea C, Kooy-Winkelaar Y, van Veelen P, de Ru A, Drijfhout JW, Koning F, et al. Fine specificity of monoclonal antibodies against celiac disease-inducing peptides in the gluteome. Am J Clin Nutr 2008;88(4):1057–66. https://doi.org/10.1093/ajcn/88.4.1057.
[12] Varma S, Krishnareddy S. Novel drug therapeutics in celiac disease: a pipeline review. Drugs 2022;82(15):1515–26. https://doi.org/10.1007/s40265-022-01784-2.
[13] Gottlieb K, Dawson J, Hussain F, Murray JA. Development of drugs for celiac disease: review of endpoints for phase 2 and 3 trials. Gastroenterol Rep 2015;3(2):91–102. https://doi.org/10.1093/gastro/gov006.
[14] Mittal P, Arora D, Parashar S, Goyal R, Khan A, Chopra H, et al. Celiac disease: pathogenesis, disease management and new insights into the herbal-based treatments. Narra J 2023;3(2):e147. https://doi.org/10.52225/narra.v3i2.147.
[15] Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov 2024;23(2):141–55. https://doi.org/10.1038/s41573-023-00832-0.
[16] Safonova A, Ghazaryan G, Stiller S, Main-Knorn M, Nendel C, Ryo M. Ten deep learning techniques to address small data problems with remote sensing. Int J Appl Earth Obs Geoinf 2023;125:103569. https://doi.org/10.1016/j.jag.2023.103569.
[17] Kwon S, Bae H, Jo J, Yoon S. Comprehensive ensemble in QSAR prediction for drug discovery. BMC Bioinf 2019;20(1):521. https://doi.org/10.1186/s12859-019-3135-4.
[18] National Center for Biotechnology Information. PubChem Protein Summary for Protein P2; 1980, Protein-Glutamine Gamma-Glutamyltransferase 2 (Human). ⟨https://pubchem.ncbi.nlm.nih.gov/protein/P21980⟩.
[19] RDKit: Open-Source Cheminformatics. ⟨https://www.rdkit.org⟩.
[20] Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Chemin 2020;12(1):43. https://doi.org/10.1186/s13321-020-00445-4.
[21] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Chemin 2015;7(1):20. https://doi.org/10.1186/s13321-015-0069-3.
[22] Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024;14(1):6086. https://doi.org/10.1038/s41598-024-56706-x.
[23] Benet LZ, Hosey CM, Ursu O, Oprea TI. BDDCS, the rule of 5 and drugability. Adv Drug Deliv Rev 2016;101:89–98. https://doi.org/10.1016/j.addr.2016.05.007.
[24] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem 2012;4(2):90–8. https://doi.org/10.1038/nchem.1243.
[25] Halder SK, Elma F. In silico identification of novel chemical compounds with Anti-TB potential for the inhibition of InhA and EthR from mycobacterium Tuberculosis. 12.04.411967. bioRxiv 2020;2020. https://doi.org/10.1101/2020.12.04.411967.
[26] Yadav R, Imran M, Dhamija P, Chaurasia DK, Handu S. Virtual screening, ADMET prediction and dynamics simulation of potential compounds targeting the main protease of SARS-CoV-2. J Biomol Struct Dyn 2021;39(17):6617–32. https://doi.org/10.1080/07391102.2020.1796812.
[27] Aykul S, Martinez-Hackert E. Determination of Half-Maximal inhibitory concentration using biosensor-based protein interaction analysis. Anal Biochem 2016;508:97–103. https://doi.org/10.1016/j.ab.2016.06.025.
[28] Mohammed A, Kora R. A comprehensive review on ensemble deep learning: opportunities and challenges. J King Saud Univ - Comput Inf Sci 2023;35(2):757–74. https://doi.org/10.1016/j.jksuci.2023.01.014.
[29] Wyner A, Olson M, Bleich J, Mease D. Explaining the success of AdaBoost and random forests as interpolating classifiers. J Mach Learn Res 2015:*18*.
[30] Chengsheng T, Huacheng L, Bing X. AdaBoost typical algorithm and its application research. MATEC Web Conf 2017;139:00222. https://doi.org/10.1051/matecconf/201713900222.
[31] Nguyen V. Bayesian optimization for accelerating hyper-parameter tuning. IEEE Second Int Conf Artif Intell Knowl Eng (AIKE) 2019;2019:302–5. https://doi.org/10.1109/AIKE.2019.00060.
[32] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Gener Hyperparameter Optim Framew 2019:2631. https://doi.org/10.1145/3292500.3330701.
[33] Altman N, Krzywinski M. Ensemble methods: bagging and random forests. Nat Methods 2017;14(10):933–4. https://doi.org/10.1038/nmeth.4438.
[34] Valdovinos RM, Sanchez JS. Ensembles of multilayer perceptron and modular neural networks for fast and accurate learning. Fifth Mex Int Conf Artif Intell 2006;2006:229–36. https://doi.org/10.1109/MICAI.2006.13.
[35] Walker AS, Pishchany G, Clardy J. Parsing molecules for drug discovery. Biochemistry 2020;59(17):1645–6. https://doi.org/10.1021/acs.biochem.0c00278.
[36] Heid E, Greenman KP, Chung Y, Li S-C, Graff DE, Vermeire FH, et al. Chemprop: a machine learning package for chemical property prediction. J Chem Inf Model 2024;64(1):9–17. https://doi.org/10.1021/acs.jcim.3c01250.
[37] Maucher IV, Rühl M, Kretschmer SBM, Hofmann B, Kühn B, Fettel J, et al. Michael acceptor containing drugs are a novel class of 5-Lipoxygenase inhibitor targeting the surface cysteines C416 and C418. Biochem Pharmacol 2017;125:55–74. https://doi.org/10.1016/j.bcp.2016.11.004.
[38] Liu Z. Michael addition reaction and its examples. Appl Comput Eng 2023;24:1–6. https://doi.org/10.54254/2755-2721/24/ojs/20230669.
[39] Jäntschi L. Molecular descriptors family on structure activity relationships 1. Review of the methodology. Leon Electron J Pract Technol 2005:*6*.
[40] Ghule, S.; Dash, S.; Bagchi, S.; Joshi, K.; Vanka, K. Predicting the Redox Potential of Phenazine Derivatives Using DFT Assisted Machine Learning. *ChemRxiv* 2022.
[41] Sarbu G, Lungu C, Balan A, Bahrin L. Synth Sulfur Contain Pipe Deriv Potential Biol Act 2014.
[42] Cao Y, Abdolmohammadi S, Ahmadi R, Issakhov A, Ebadi AG, Vessally E. Direct synthesis of sulfenamides, sulfinamides, and sulfonamides from thiols and amines. RSC Adv 2021;11(51):32394–407. https://doi.org/10.1039/D1RA04368D.