International Journal of
Molecular Sciences

MDPI

Article

Machine Learning-Driven Classification of Urease Inhibitors Leveraging Physicochemical Properties as Effective Filter Criteria

Natalia Morales [1], Elizabeth Valdés-Muñoz [2], Jaime González [1], Paulina Valenzuela-Hormazábal [3], Jonathan M. Palma [4], Christian Galarza [5], Ángel Catagua-González [5], Osvaldo Yáñez [6], Alfredo Pereira [7,*] and Daniel Bustos [8,*]

1 Magíster en Ciencias de la Computación, Universidad Católica del Maule, Talca 3460000, Chile; nmoralesr@ucm.cl (N.M.); jaime.gonzalez@alu.ucm.cl (J.G.)
2 Doctorado en Biotecnología Traslacional, Centro de Biotecnología de los Recursos Naturales, Universidad Católica del Maule, Talca 3480094, Chile; elizabeth.valdes@alu.ucm.cl
3 Departamento de Farmacología, Facultad de Ciencias Biológicas, Universidad de Concepción, Concepción 4030000, Chile; paulinvalenzuela@udec.cl
4 Facultad de Ingeniería, Universidad de Talca, Curicó 3344158, Chile; jonathan.palma@utalca.cl
5 Departamento de Matemáticas, Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, Guayaquil EC090903, Ecuador; chedgala@espol.edu.ec (C.G.); anglucat@espol.edu.ec (Á.C.-G.)
6 Núcleo de Investigación en Data Science, Facultad de Ingeniería y Negocios, Universidad de las Américas, Santiago 7500000, Chile; oyanez@udla.cl
7 Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Bellavista 7, Santiago 8420524, Chile
8 Laboratorio de Bioinformática y Química Computacional, Departamento de Medicina Traslacional, Facultad de Medicina, Universidad Católica del Maule, Talca 3480094, Chile
* Correspondence: alfredo.pereira@uss.cl (A.P.); dbustos@ucm.cl (D.B.)

check for updates

Citation: Morales, N.; Valdés-Muñoz, E.; González, J.; Valenzuela-Hormazábal, P.; Palma, J.M.; Galarza, C.; Catagua-González, Á.; Yáñez, O.; Pereira, A.; Bustos, D. Machine Learning-Driven Classification of Urease Inhibitors Leveraging Physicochemical Properties as Effective Filter Criteria. *Int. J. Mol. Sci.* **2024**, *25*, 4303. https://doi.org/10.3390/ijms25084303

Academic Editor: Dong-Jun Yu

Received: 15 March 2024
Revised: 3 April 2024
Accepted: 8 April 2024
Published: 13 April 2024

Abstract: Urease, a pivotal enzyme in nitrogen metabolism, plays a crucial role in various microorganisms, including the pathogenic *Helicobacter pylori*. Inhibiting urease activity offers a promising approach to combating infections and associated ailments, such as chronic kidney diseases and gastric cancer. However, identifying potent urease inhibitors remains challenging due to resistance issues that hinder traditional approaches. Recently, machine learning (ML)-based models have demonstrated the ability to predict the bioactivity of molecules rapidly and effectively. In this study, we present ML models designed to predict urease inhibitors by leveraging essential physicochemical properties. The methodological approach involved constructing a dataset of urease inhibitors through an extensive literature search. Subsequently, these inhibitors were characterized based on physicochemical properties calculations. An exploratory data analysis was then conducted to identify and analyze critical features. Ultimately, 252 classification models were trained, utilizing a combination of seven ML algorithms, three attribute selection methods, and six different strategies for categorizing inhibitory activity. The investigation unveiled discernible trends distinguishing urease inhibitors from non-inhibitors. This differentiation enabled the identification of essential features that are crucial for precise classification. Through a comprehensive comparison of ML algorithms, tree-based methods like random forest, decision tree, and XGBoost exhibited superior performance. Additionally, incorporating the "chemical family type" attribute significantly enhanced model accuracy. Strategies involving a gray-zone categorization demonstrated marked improvements in predictive precision. This research underscores the transformative potential of ML in predicting urease inhibitors. The meticulous methodology outlined herein offers actionable insights for developing robust predictive models within biochemical systems.

Keywords: urease inhibitors; cheminformatics; machine learning; predictive modeling; bioactivity prediction; classification models

1. Introduction

1.1. Urease Enzyme and Its Implications in the Human Context

The urease enzyme is a key element in nitrogen (N) metabolism in bacteria, fungi, algae, and plants, hydrolyzing urea (carbamide) over 10^{14}-fold the conversion rate to ammonia and CO_2 [1,2]. One of the most conflictive ureolytic bacteria for humans is *Helicobacter pylori* (H.p.), which is very well adapted to survive in a wide range of environments by secreting a high number of urease enzymes, even in acidic environments such as the stomach [3,4]. It is estimated that over 50% of the world population is already infected by H.p. [5,6]. Once H.p. colonizes the host, this Gram-negative bacterium increases the risk for peptic ulcers [7,8], chronic kidney diseases [9], idiopathic thrombocytopenic purpura [10,11], iron deficiency anemia [12], and gastric cancer (GC) such as gastric adenocarcinoma [13,14] and MALT (mucosa-associated lymphoid tissue) lymphoma [15–17]. H.p. is the only bacterium classified in Group I of carcinogens to humans by the International Agency for Research on Cancer, where 89% of all GC is related to H.p. infections [18]. Until 2020, GC was one of the deadliest types of cancer worldwide [19]. In this sense, it has been evidenced that H.p. eradication reduces the mortality rate caused by GC [20]. The current treatment proposed to eradicate H.p. infections is to combine broadband antibiotics (amoxicillin, metronidazole, and clarithromycin) with a proton-pump inhibitor or with bismuth-containing compounds [21–23]. Nowadays, these therapies have become unfeasible due to both the alarming resistance of H.p. to antibiotics worldwide and the side effects (nausea, diarrhea, headache, angioedema, and microflora disorders) produced by antibiotics. New therapeutical strategies based on the use of urease inhibitors (UIs) have been proposed to treat infections of urease-dependent microorganisms such as H.p since inhibiting the ureolytic faculty in H.p. causes this bacterium to become unable to cause infections in animal models [24]. However, available classic UIs (sulfhydryl compounds, amides and esters of phosphoric acid, hydroxamic acid derivatives, and imidazoles) are toxic for humans, which precludes their clinical uses [25].

Many efforts are being focused on searching for potent UIs [26]. Natural sources provide an immeasurable number of organic compounds with anti-urease activity, and these compounds could be used as a starting point to design stronger inhibitors. At present, there exist compounds reported with anti-urease activity such as polyphenols, flavonoids [27–29], alkaloids [30], triazole [31–35] thiadiazole [24,30,31,36], and coumarins [24,37–41]. Despite the existence of a significant repertoire of urease inhibitors, there are compelling reasons to continue the pursuit of novel inhibitors. This drive is fueled by the limitations of conventional methods and the potential advantages offered by unconventional techniques, such as ML. These ML models can provide insights into the structure–activity relationships of urease inhibitors, aiding in the rational design of novel compounds while reducing time and resources.

1.2. Machine Learning in the New Era of Computer-Aided Drug Discovery

It is projected that bringing a drug from its initial stage to market may take up about a decade and incur expenses exceeding USD 2.8 billion [42]. The early stage of drug discovery, also named computer-aided drug discovery (CADD), has rapidly emerged along with the development of structural biology and the computational power of new hardware [43]. CADD is a collection of diverse computational techniques and resources, comprising compound databases, molecular simulations, structure- and ligand-based virtual screenings (VS), hit and lead optimization, quantitative structure–activity relationship (QSAR), among many others. The integration of various ML algorithms into the CADD process has greatly benefited pharmaceutical companies and academic research as ML provides them with innovative and efficient ways in every stage of the CADD process [44,45] and other branches of chemistry [46–48]. In ML, there exist two primary categories of algorithms: supervised learning and unsupervised learning. The first is responsible for learning from labeled training samples to determine the labels of new samples, whereas the latter is responsible for identifying patterns within an unlabeled dataset. Typically, before pattern recognition,

high-dimensional data are transformed into a lower dimension via unsupervised learning algorithms to increase efficiency [49]. Through the use of ML algorithms, various models have been created which allow for a more precise understanding of the biochemical and physical–chemical characteristics of candidate compounds in VS protocols, thus allowing a reduction in false positives and false negatives [50]. In this sense, Liu and colleagues utilized a support vector machine (SVM) algorithm to construct various classification models for 85 specific cyclooxygenase-2 inhibitors featuring the 1,5-diarylimidazoles scaffold. The optimal classification models show accuracies of 91.2% and 88.2% for the training and test sets, respectively [51]. Kumar and Patra employed known catechol O-methyltransferase inhibitors as input to discover new inhibitors by combining ML regression and sampling molecular dynamics methods. Their models achieved an R^2 of over 0.7 for both training and test datasets [52]. In order to deal with the imbalance of inhibitor and non-inhibitor classes, Tinivella et al. employed a flexible thresholding strategy on a set of modulators deposited in ChEMBL of two human carbonic anhydrase isoforms through an ML protocol [53]. In the urease context, to our best knowledge, only two studies have employed ML techniques for the development of new inhibitors. Mermer et al. [54] uses regression and classification models to identify novel thiazole derivatives with a balanced accuracy of 78% and an R^2 between 0.2 and 0.7. Aniceto et al. [55] discover new inhibitors of *jack bean* urease by using three ML algorithms with 81% precision in its best model.

The main goal of this research is to construct robust and accurate ML models capable of predicting the activity of urease inhibitors in *Hp* (see schematic workflow in Supplementary Figure S1). Simultaneously, we conduct a thorough exploratory data analysis to identify pertinent features crucial for predicting the behavior of urease inhibitors. Subsequently, our objective is to investigate diverse strategies for categorizing bioactivity and identify suitable ML algorithms to determine the most effective approach for model development. The study encompasses the standard procedures in ML protocols, which include data collection, data preprocessing, exploratory data analysis (EDA), data partitioning, and the learning stages (model selection, training, hyperparameter tuning, and evaluation).

2. Results and Discussion

2.1. Exploratory Data Analysis

During the analysis of this dataset, we investigated 677 different compounds and 207 variables, with a specific focus on the response variable IC_{50}. Opting to utilize pIC_{50} instead of IC_{50} proves beneficial due to the notable variation in concentration ranges exhibited by the latter (ranging from 0.009 μM to 1000 μM within our dataset). IC_{50} values frequently encompass a wide span of orders of magnitude, consequently giving rise to challenges in direct comparison and effective visualization. In contrast, pIC_{50} offers a more concise representation achieved by applying the negative logarithm to IC_{50} values, which are then standardized to a consistent concentration level. This transformation serves a dual purpose: not only does it normalize the data, but it also enhances the comprehension of compound potency across a wider array of concentrations. In Supplementary Figure S2, the distribution of bioactivity is illustrated, delineating nUIs, UIs, and the intermediate gray-zone compounds generated by the 5–50 μM cutoff.

To identify potential inhibitor features, a correlation analysis is performed with respect to pIC_{50}, which substantially increases correlations compared to IC_{50}. Figure 1 displays a heatmap of the correlation matrix between pIC_{50} and the 16 features with $|\rho| > 0.4$. Larger circle diameters indicate stronger correlations. WTPT.4, TPSA, and RPSA exhibit moderate and negative correlations with pIC_{50}. These attributes pertain to topological (WTPT.4) and electronic (TPSA and RPSA) characteristics. The initial one, WTPT.4, signifies the molecular branching in each molecule originating from oxygen atoms. On the other hand, TPSA and RPSA correspond to diverse measurements of the solvent-accessible surface area (SASA). Specifically, TPSA represents the sum of the SASA of atoms with an absolute value of partial charges ≥ 0.2, while RPSA is calculated as the ratio of TPSA to the total

SASA. Additionality, positive relationships are observed with khs.dsN (Kier–Hall E-state descriptor), C1SP2 (carbon atoms with hybridization Sp2), SCH.6 (Kier and Hall Chi chain index), WTPT.5 (molecular branching starting from the nitrogen), MDEN.22, and MDEN.23 (molecular distance edge descriptors). Significant and strong positive correlations among the independent variables are evident, such as TPSA and RPSA (0.92), FNSA.3 and RHSA (0.94), khs.aasN and MDEN.23 (0.95), as well as RnRings5 and SCH.5 (0.99), among several others. These correlations are unsurprising, given that these descriptors belong to the same categories and are closely interconnected. Given that none of the selected features exhibit a high correlation with the response variable pIC_{50}, it suggests a lack of strong linear association between the chemical properties of the compounds and their inhibitory activity against urease. Consequently, the molecular descriptors captured by these variables may not directly or simply correspond to the bioactivity measured by pIC_{50}. This circumstance prompts the exploration of supervised machine learning models as they can adeptly capture complex and non-linear relationships between the features and pIC_{50} by discerning non-trivial chemical patterns within our dataset.

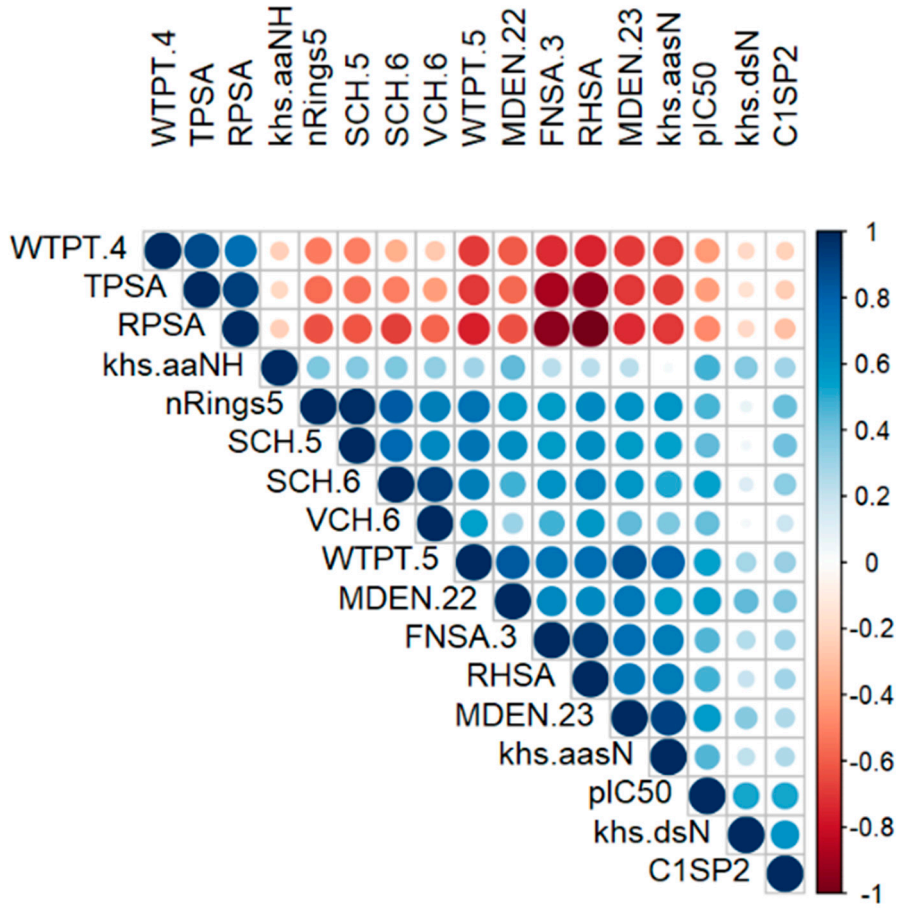

Figure 1. Heatmap of the correlation matrix. The heatmap was plotted by considering the negative logarithm of response variable (pIC50) and the most strongly correlated features ($\rho > 0.4$). The size of the circle, as well as the color, reflects the intensity of the correlation of the two variables found at the intersection of the matrix. The blue color reflects positive correlations, while the red color reflects negative correlations. Finally, white reflects an absence of correlation.

The central tendency (median) and variability (interquartile range) measures were studied for the UI and nUI groups with the most correlated features. The boxplots in Figure 2 show clear differences in the 16 relevant features between UI and nUI compounds. The median and variability differ between the two groups, and outliers are present in most features. Features with lower median values for inhibitor compounds include WTPT.4,

TPSA, and RPSA, while the remaining features have higher median values allowing separate both classes. Furthermore, WTPT.4, TPSA, and RPSA not only have lower median values for the inhibitor group but also lower variability, resulting in more consistent values and lower uncertainty. Finally, all 16 features show statistically significant mean differences between the two groups based on the *p*-value of a joint Wilcoxon rank sum test.

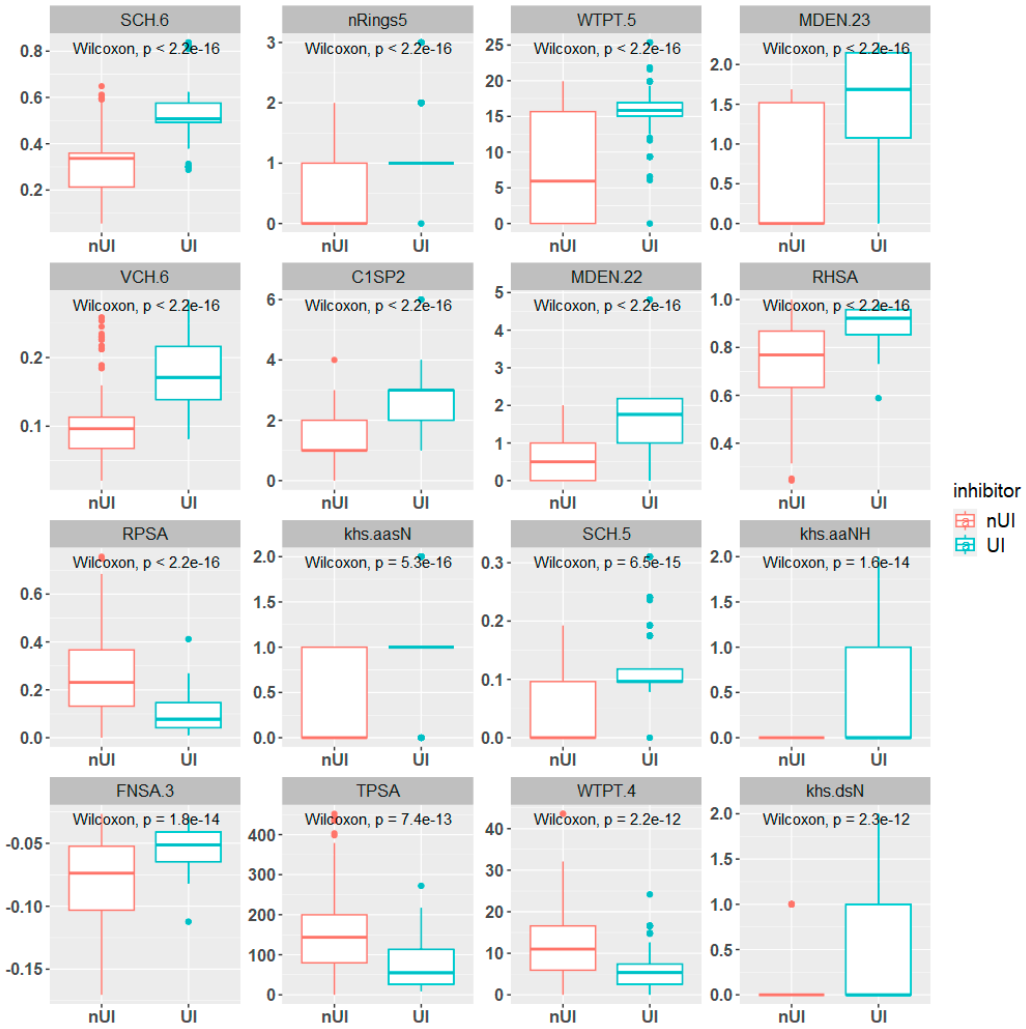

Figure 2. Boxplots for the 16 most relevant features. Each feature plotted was separated according to the response variable using the double cutoff 5 μM and 50 μM. The *p*-value was computed with a statistical Wilcoxon rank sum test.

Figure 3 presents a two-dimensional representation of the data using a PCA. The direction of each arrow represents the direction and magnitude of the maximum variability in the data in the chemical feature space. Each arrow, also known as an eigenvector, indicates the direction in which the data have the greatest variation: the larger the size, the greater the variation. The ellipsoids (red for nUIs and blue for UIs) represent the spread of the data for each class in the feature space reduced by the method. Each ellipsoid describes the cloud of points of a class in the lower-dimensional feature space generated by the PCA. The shape and size of the ellipsoids indicate the variability in and distribution of the data for each class. It is revealed from this visualization that certain features, such as khs.aaNH, SCH.5, SCH.6, C1SP2, VCH.6, and nRings5, are associated with UIs, while TPSA, WTPT.4, and RPSA values are associated with nUIs. The intersection between the ellipsoids represents the region where the two classes have an overlap in the reduced feature space. This overlap indicates that there are data instances that share similar characteristics between

both classes. Therefore, the intersection between the ellipsoids may contain points that are difficult to definitively classify as belonging to a specific class, suggesting the presence of instances with ambiguous or intermediate bioactivity. It is mainly on these that we hope that later ML methods will allow them to be classified correctly. Moreover, we conduct a t-SNE analysis, as showcased in Supplementary Figure S3. t-SNE reveals a separation between UI and nUI classes where compounds situated within the gray zone exhibit overlapping representations in both clusters (UIs and nUIs). This highlights, together with the results of the PCA, the essential requirement to explore more intricate bioactivity categorization approaches, with the aim of precisely distinguishing which compounds can be appropriately classified as UIs. The exploratory analysis conducted underscores the presence of a diverse range of features that facilitate clear distinctions both within individual classes and among different groups. This forms a robust basis for crafting ML classification models, which have the potential to unveil complex and less obvious relationships during the preliminary stages of EDA.

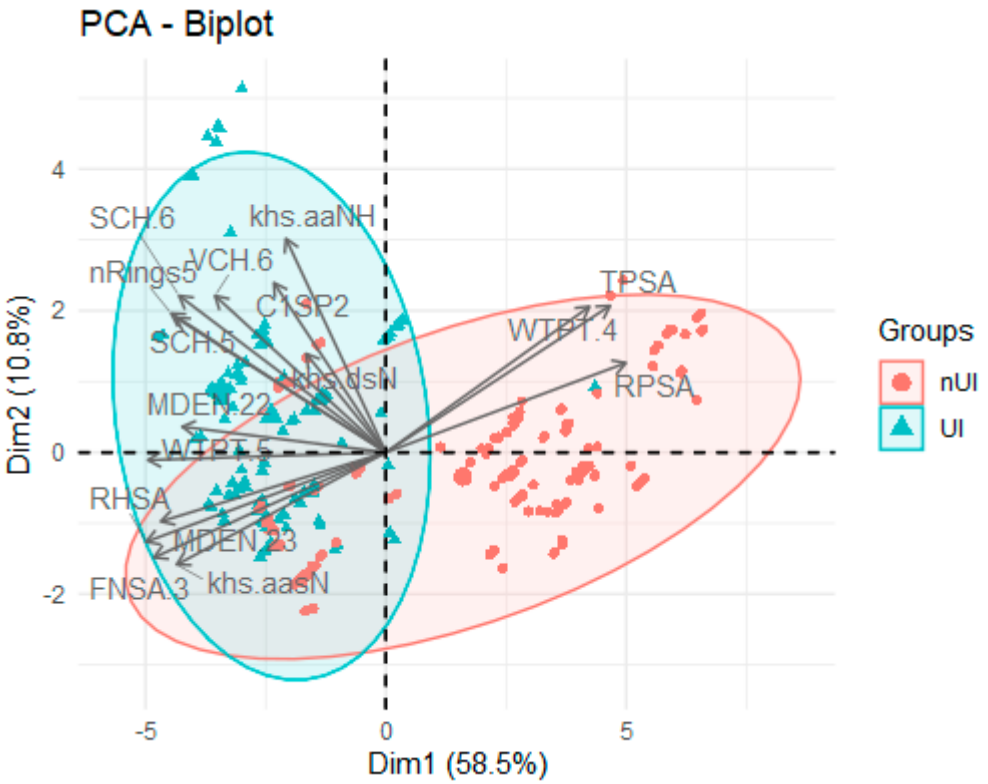

Figure 3. Two-dimensional representation of the data and relevant features using PCA where each figure represents a distinct compound categorized as a urease inhibitor (UI) indicated in cyan, encompassing molecules exhibiting an $IC_{50} \leq 5$ μM. Similarly, non-urease inhibitors (nUIs) are denoted in red, encompassing molecules with an $IC_{50} \geq 50$ μM. Additionally are showed the features contributing to each class.

2.2. Machine Learning Models

In the present study, an extensive methodology was employed to discover high-performance models for the prediction of urease inhibitors. Seven ML algorithms, random forest (RF), support vector machine (SVM), decision tree (DT), eXtreme Gradient Boosting (XGB), k-nearest neighbor (KNN), naive Bayes (NB), and logistic regression, (LR) and three attribute selection methods, Boruta, XGB, and nFS (non-feature selection), were compared in conjunction with six different strategies for categorizing the bioactivity of the inhibitors, resulting in a total of 126 models. Furthermore, each model was trained twice, considering the attribute "chemical family type" and excluding it. In summary, 252 distinct models were trained.

Figure 4 displays a comparison of the seven ML algorithms. It can be observed that each algorithm was combined with the three attribute selection methods, and each one was executed separately, considering (Figure 4A) or excluding (Figure 4B) the "chemical family type" attribute. It is crucial to mention that the six MCC values used to construct each boxplot correspond to the six bioactivity categorization strategies (Table 1). As a result, the best and worst models are labeled using the categorization cutoffs, while the best and worst average algorithms are indicated by the red and black arrows, respectively. Based on these findings, noticeable differences among the various algorithms become evident. The algorithm with the poorest average performance is NB, regardless of whether the chemical family type is considered or not. However, the individual model with the worst MCC performance occurs when LR is used in combination with the bioactivity categorization based on a 5 µM cutoff and without utilizing an attribute selection method, yielding only a 0.25 MCC score. On the other hand, the algorithms with the best average performance are DT, when the chemical family type is not considered, and XGB, when it is considered. Both algorithms employ XGB as the attribute selection method. Regarding individual models that do not consider the chemical family type, the best is RF, using the BORUTA attribute selection method and combined with the bioactivity categorization based on 10–50 µM cutoffs (RF_BORUTA_10–50), achieving a 0.84 MCC score. The top individual models that consider the chemical family type are DT, using the BORUTA attribute selection method and combined with the bioactivity categorization based on 5–50 µM cutoffs (DT_BORUTA_5–50), achieving a 0.97 MCC score. Additionally, XGB, without an attribute selection method and combined with the bioactivity categorization based on a 5 µM cutoff (XGB_nFS_5), achieves a 0.97 MCC score.

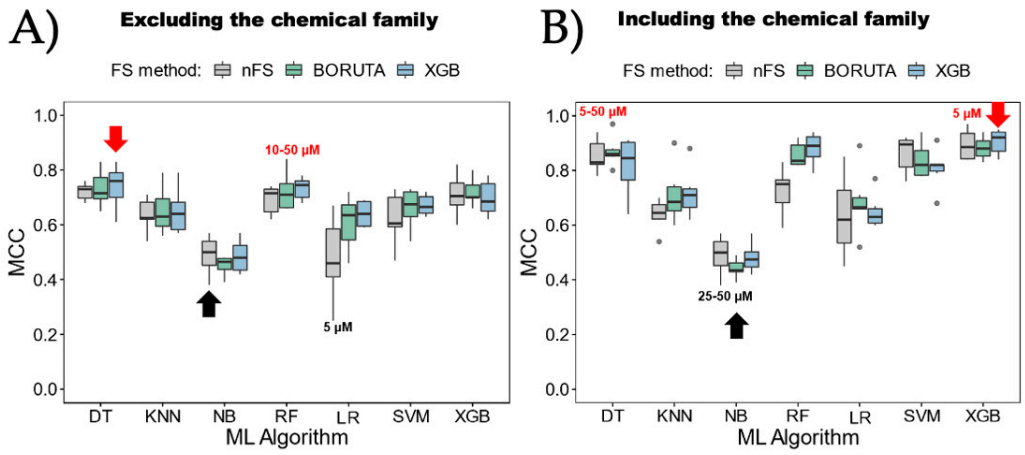

Figure 4. Comparison of ML algorithms. Each algorithm was compared through Matthews correlation coefficient (MCC) (**A**) excluding the chemical family and (**B**) including it. Black and red labels indicate lowest and highest MCC values individually per strategy for bioactive characterization. Meanwhile, black and red arrows show the lowest and highest MCC values as averages considering all the strategies tested in each algorithm. nFS: non-feature selection; DT: decision tree; KNN: k-nearest neighbor; RF: random forest; LR: logistic regression; SVM: support vector machine; XGB: eXtreme Gradient Boosting. The points at the ends of the boxplots show the outliers in each comparison. Calculated from those values that are below: Q1 − 1.5 ∗ IQR or above Q3 + 1.5 ∗ IQR. IQR being the Interquartile Range.

Table 1. Proposed prediction strategies based on the IC_{50} values for bioactive characterization.

One-Cutoff Strategies	Two-Cutoff Strategies
UIs < 5 µM < nUIs	UIs < 5 µM < gray zone < 50 µM < nUIs
UIs < 10 µM < nUIs	UIs < 10 µM < gray zone < 50 µM < nUIs
UIs < 25 µM < nUIs	UIs < 25 µM < gray zone < 50 µM < nUIs

A comparison to identify the optimal bioactivity categorization strategy for inhibitors is presented in Figure 5. In this case, each strategy is combined with the three attribute selection methods and executed separately, including or excluding the "chemical family type" attribute. It is important to note that the seven MCC values used to construct each boxplot correspond to the seven ML algorithms; hence, the best and worst models are labeled with the algorithm names, while the best and worst average strategies are indicated by the red and black arrows, respectively. Analyzing these results, the differences are not as pronounced as in the comparison of ML algorithms. However, it is observed that the best average strategy is when using the 5–50 μM cutoffs in combination with XGB as the attribute selection method. This holds true for both cases, including or excluding the "chemical family type" attribute. Since the data used to construct Figures 4 and 5 (MCC scores of the models) are the same, the models with the best and worst performance coincide between both figures. Thus, as can be observed, RF_BORUTA_10–50 emerges as the best model when the "chemical family type" attribute is excluded, while DT_BORUTA_5–50 and XGB_nFS_5 are the best models when the "chemical family type" attribute is included.

The data collection process, as described in the methodology, underscores the robustness of our study. Specifically, the calculation of molecular descriptors using the rCDK package generated a comprehensive set of 290 parameters across five categories: 'topological', 'electronic', 'constitutional', 'hybrid', and 'geometrical'. From these initial parameters, 83 were excluded due to their high variability and minimal contribution to information. The final set of 207 descriptors, along with an extra categorical attribute representing the chemical family, were meticulously explored even with Boruta and XGB as feature selection methods to ensure their relevance to physicochemical coupling with the urease binding site. This rigorous selection process aimed to enhance the predictive capabilities of our models and provide valuable insights into potential drug discovery pathways. However, there was not a clear preference for one ML algorithm over another, even with better MCC performances without feature selection (considering the 207/208 descriptors). It is also important to notate that the characteristics selected in both the supervised and unsupervised processes coincide to a large extent (Supplementary file: "Features_selected.xlsx"), supporting that these characteristics were a good input for the construction of classification models.

The effectiveness of tree-based methods such as RF, DT, and XGB can be attributed to several factors. Firstly, these algorithms are capable of capturing non-linear relationships and interactions between features, which are often present in complex biological datasets. Additionally, tree-based models inherently handle feature importance, allowing for the identification of key molecular descriptors contributing to bioactivity prediction. Moreover, ensemble methods like random forest and XGB further enhance predictive performance by aggregating multiple decision trees, thereby reducing overfitting and improving generalization to unseen data.

Finally, Figure 6 presents a direct comparison of the top four models, whether including the "chemical family type" attribute or not. First and foremost, it is observed that all models achieve an AUC greater than 0.93. However, the best performances are attained when the "chemical family type" of the inhibitors is considered as an attribute. As mentioned previously, the best model when excluding the "chemical family type" attribute is RF_BORUTA_10–50, which achieves an AUC of 0.9928 in this analysis. On the other hand, when the "chemical family type" attribute is considered, the top identified models are DT_BORUTA_5–50 and XGB_nFS_5, both of which achieve a perfect AUC of 1, indicating flawless classification between inhibitors and non-inhibitors. Another crucial aspect to mention regarding the algorithms is that all eight models presented in Figure 6 are tree-based methods, surpassing other models like SVM, LR, KNN, or NB. From the perspective of attribute selection methods, there seems to be no direct preference for one over the other; therefore, all of them could be viable for implementing tree-based models. As for the categorization strategies, there also appears to be no clear preference when analyzing these ROC curves. However, considering the results from Figure 5, it is inferred that the best outcomes are obtained when using the strategies with a gray zone,

meaning the use of two cutoffs to categorize UIs and nUIs. The ROC curves emphasize their ability to distinguish between UIs and nUIs. Achieving high AUC values in ROC curves is crucial in drug discovery as it reflects the model's ability to correctly classify compounds into their respective categories. High AUC values indicate strong predictive performance, suggesting that the models are capable of accurately identifying potential UIs. In a clinical context, these models could play a vital role in accelerating the drug discovery process by prioritizing compounds with a higher likelihood of urease inhibition for further experimental validation. Additionally, the biological relevance of the models' predictive accuracy underscores their practical utility. The accurate prediction of bioactivity enables researchers to focus resources on compounds with the greatest potential for therapeutic intervention, thereby facilitating the development of novel treatments for conditions such as *Hp* infection.

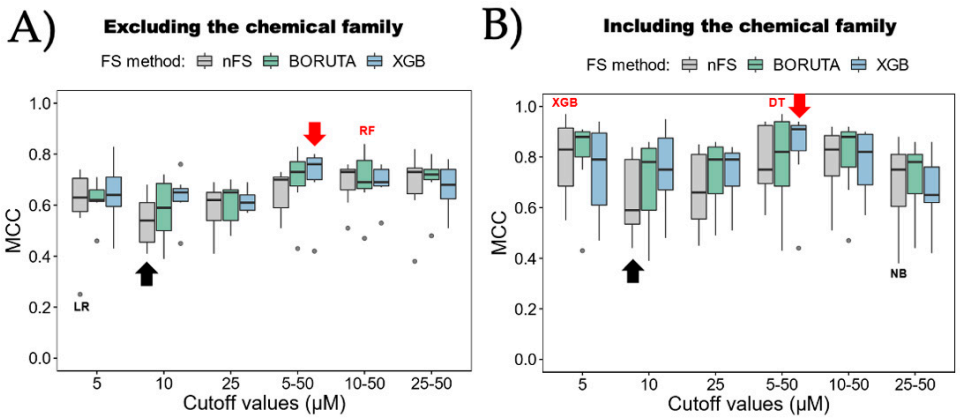

Figure 5. Comparison of strategies for bioactivity categorization. Each strategy was compared through Matthews correlation coefficient (MCC) (**A**) excluding the chemical family and (**B**) including it. Black and red labels indicate lowest and highest MCC values individually per ML algorithm. Meanwhile, black and red arrows show the lowest and highest MCC values as averages considering all the ML algorithm tested in that strategy. nFS: non-feature selection. The points at the ends of the boxplots show the outliers in each comparison. Calculated from those values that are below: Q1 − 1.5 ∗ IQR or above Q3 + 1.5 ∗ IQR. IQR being the Interquartile Range.

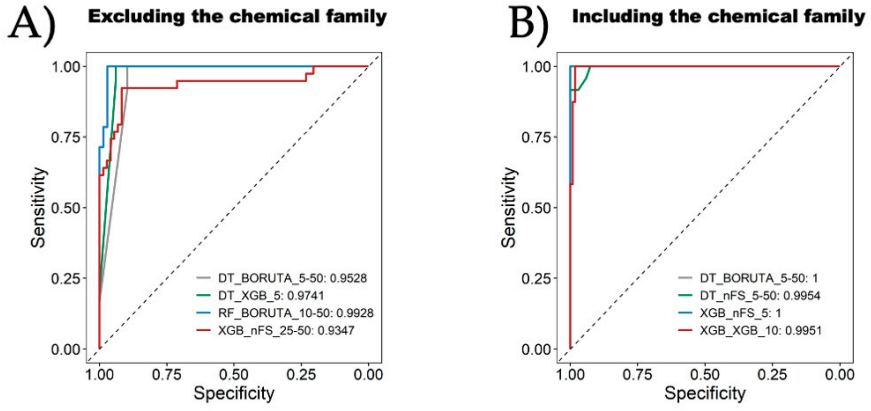

Figure 6. Comparison of the best four models (**A**) excluding the chemical family and (**B**) including it. The legend shows the algorithm followed by the feature selection and the strategy for bioactive characterization associated with the area under the curve (AUC) for each model.

3. Materials and Methods

3.1. Data Collection

A scientific literature exploration was carried out on the Web of Science (WOS) database, utilizing the search terms "urease inhibitors" AND "*Helicobacter pylori*". The

search was limited to articles published from 2010 onwards. Afterward, a categorization of inhibitors was produced by grouping them according to their chemical family. We found compounds belonging to flavonoids [27–29], alkaloids [30], triazole [31–35], thiadiazole [24,30,31,36], and coumarins [24,37–41] chemical families. This dataset involved an IC_{50} range from 0.009 micromolar (μM) to concentrations where a minimum inhibitory concentration for *Hp* urease enzyme was not determined, here termed non-urease inhibitors (nUIs). This is because some molecules were compounds whose inhibitory concentration was not detected in the experiments; in other words, they were molecules without inhibitory potency for *Hp* urease. Therefore, they did not have a numerical value associated with the response variable. Thus, considering the biological criteria and empirical values of non-inhibitory compounds reported in the literature, an arbitrary IC_{50} value of 1 mM was assigned to all these molecules. Lastly, those compounds whose inhibition measurement was not carried out by calculating the half-maximal inhibitory concentration (IC_{50}) were discarded. The UIs and nUIs previously collected were drawn utilizing the 2D Sketcher tool from the Maestro Schrodinger suite [56], which was also used to add their corresponding valences. To assign protonation states, the Epik tool [57] was utilized at a pH of 7.2, which is the standard pH at which biological assays are typically carried out in urease. Finally, each of the 667 molecules were converted to an SDF format for subsequent analysis.

3.2. Characterization and Preprocessing

The calculation of molecular descriptors was performed using the rCDK package version 3.6.0. [58] from the Chemistry Development Kit library in the R programming environment [59]. All available categories of descriptors in this library ("topological", "electronic", "constitutional", "hybrid", and "geometrical") were calculated, generating a total of 290 descriptors (Table S1 in Supplementary Information). Subsequently, data processing was carried out to discard any variable with minimal information contribution and/or to impute missing data in specific variables. To do this, firstly, a criterion was generated to exclude attributes that had 80% or more information loss. As a result of this, 286 descriptors remained, and the excluded variables were Wgamma1.unity, Wgamma2.unity, Wgamma3.unity, and WG.unity. Then, using the multiple imputation by chained equations method [60] with 3 iterations and 3 imputations, missing data were completed for the variables Weta1.unity and WD.unity. Next, variables with variance close to zero were excluded using the nearZeroVar function of the Caret package version 6.0-94 [61], resulting in a total of 207 molecular descriptors. It is important to mention that, in addition to the 207 descriptors, an extra categorical attribute corresponding to the type of chemical family previously recorded was considered. This attribute had significant relevance for the subsequent stages of the study as it may or may not have been included in the models according to the strategy used.

3.3. Exploratory Data Analysis

The dataset comprised 677 examples (compounds), 207 numerical variables (molecular descriptors), and 1 categorial variable (family type), and the response variable (IC_{50} in μM) was subjected to an EDA with the double aim of (1) detecting the existence of a correlation between descriptors studied and (2) identifying whether all descriptors or a subset of them enabled a clear differentiation between the UI and nUI classes in the response variable. Furthermore, IC_{50} is not a linear measure, and hence, it does not allow the adequate separation of the classes. We transformed the variable response into pIC_{50}, the negative logarithm of IC_{50}. This transformation is also commonly applied in statistical contexts to positive quantities to symmetrize data. On the other hand, and as mentioned before, this work aimed to predict UIs through binary classification models. For this purpose, we assigned a cutoff in the response variable to maximize the separation between the UI and nUI classes, where UIs had an $IC_{50} \leq 5$ μM ($pIC_{50} \leq 5.30$) and nUIs had an $IC_{50} \geq 50$ μM ($pIC_{50} \geq 4.30$). For the 207 numerical descriptors, we analyzed the correlation between variables through a heatmap of the correlation matrix by using the corrplot package in R.

We considered variables with a correlation magnitude (regardless of the sign) of >0.4. The variables exhibiting the highest correlations with the response variable were considered the most promising candidates for constructing predictive models. This stemmed from the fact that alterations in these highly correlated variables tend to correspond with shifts in the response variable. Therefore, the correlated features were used (1) as input to analyze the central tendency (median) and variability (interquartile range) measures in UI and nUI classes with a statistical Wilcoxon rank sum test and (2) to separate both classes through a principal component analysis (PCA) using the built-in R functions prcomp(). Additionally, we visualized the chemical space of the dataset by using the Python library ChemPlot [62,63] through a t-distributed stochastic neighbor embedding (t-SNE) analysis with 1000 iterations and perplexity = 30. t-SNE is a non-linear dimensionality reduction technique that is particularly effective at preserving the local structure of the data. Here, we presented the dimensionality of UIs, nUIs, and those molecules that did not fit into the predefined classes, here called the gray zone.

3.4. Strategies for Bioactivity Categorization (Data Splitting)

In the previous EDA, we employed a cutoff to separate and categorize both classes. Particularly, the t-SNE analysis indicated that the chosen cutoff (5–50 μM) effectively separated the classes. However, the presence of gray points representing the gray zone highlighted the need for deeper exploration. These data points, with their ambiguous bioactivity, required further scrutiny to optimize the class separation while minimizing data loss. In this sense, in the ML scheme, more than one cutoff concentration was used for this bioactivity categorization, giving rise to different classification tasks and at the same time different strategies to predict UIs. The details of the cutoffs used, and the proposed prediction strategies, are presented in Table 1. Six strategies were planned, three based on 1 cutoff and three based on 2 cutoffs. Basically, when there were 2 cutoffs, compounds with IC_{50} values greater than cutoff 1 and lower than cutoff 2 (compounds in the gray zone) were excluded from the training and testing of the models. Instead, in the case of employing a single cutoff, no compound was excluded within the ambiguous gray zone, thus defining compounds under the cutoff as UIs and those surpassing it as nUIs. The quantities of UIs and nUIs for each strategy are presented in Tables 2 and 3. It is important to note that the use of either IC_{50} or pIC_{50} is irrelevant for ML models as they can naturally model non-linearities through various non-linear transformations during preprocessing.

Table 2. UIs and nUIs for 1-cutoff strategies for bioactive characterization.

	IC_{50}: 5 μM		**IC_{50}: 10 μM**		**IC_{50}: 25 μM**	
	N°	**%**	**N°**	**%**	**N°**	**%**
UIs	119	18	145	22	221	33
nUIs	558	82	532	78	456	67

Table 3. UIs and nUIs for 2-cutoff strategies for bioactive characterization.

	IC_{50}: 5 μM and 50 μM		**IC_{50}: 10 μM and 50 μM**		**IC_{50}: 25 μM and 50 μM**	
	N°	**%**	**N°**	**%**	**N°**	**%**
UIs	119	26	145	30	221	39
nUIs	341	74	341	70	341	61

3.5. Training and Testing of Inhibitory Classification Models

In the six proposed strategies, the data were distributed between training and testing in an 80:20 ratio (Table 4). Seven supervised ML algorithms were used and compared to train the models: RF, SVM, DT, XGB, KNN, NB, and LR. In addition, Boruta [64] and XGB [65] were used separately as feature selection methods during training to compare them with the

models built using all the attributes (nFS). Boruta utilizes a random forest-based approach to identify relevant features by comparing their importance with randomized counterparts. On the other hand, XGB assesses feature importance by training multiple decision trees and evaluating their frequency of use in decision-making processes. Repeated cross-validation (10 folds and 5 repetitions) was used to train and validate the models. Furthermore, SMOTE [66] was applied to balance the classes in each training step. Moreover, the hyperparameters of each algorithm were optimized during cross-validation using a grid search method. The models were both trained and tested with consideration for the chemical family of the compounds within the dataset and without taking this parameter into account. All ML algorithms, feature selection methods, SMOTE, and cross-validations were executed in R using the Caret package functions. Finally, to evaluate the models, the Matthews correlation coefficient (MCC) and the area under the ROC curves (AUC-ROC) were calculated.

Table 4. Data distribution between training and testing per strategy.

1-Cutoff Strategies			2-Cutoff Strategies		
IC_{50}	**Training**	**Testing**	**IC_{50}**	**Training**	**Testing**
IC_{50}: 5 μM	542	135	IC_{50}: 5 μM and 50 μM	368	92
IC_{50}: 10 μM	542	135	IC_{50}: 10 μM and 50 μM	389	97
IC_{50}: 25 μM	542	135	IC_{50}: 25 μM and 50 μM	450	112

4. Conclusions

In conclusion, this study provides valuable insights into the prediction of urease inhibitors using cheminformatics and ML approaches. Through a comprehensive methodology and rigorous analysis, several key conclusions can be drawn:

1. Algorithm Preference: The study recommends favoring tree-based methods, including random forest (RF), decision tree (DT), and eXtreme Gradient Boosting (XGB), over other algorithms like k-nearest neighbor (KNN), support vector machine (SVM), naive Bayes (NB), or logistic regression (LR) for inhibitor classification.
2. Attribute Selection Influence: While attribute selection methods could potentially improve model performance, their influence varies based on the ML algorithm chosen. There is not a clear preference for one method over another, suggesting their implementation should be algorithm-specific.
3. Effective Categorization Strategies: The exploratory data analysis and ML analysis recommend employing strategies that involve a gray zone, utilizing two cutoffs for categorizing urease UIs and nUIs. These strategies tend to yield better model performance, offering improved accuracy in classification tasks, reaching almost 10 percent over one-cutoff strategies in our models. By delineating these boundaries, we can effectively train our models to distinguish between active and inactive compounds, thus enhancing the accuracy of our predictions. Moreover, understanding the implications of these cutoffs is critical for optimizing model performance. Nevertheless, strategies with a gray zone can lead to better performance, and it is crucial to consider the biological implications. Expanding the gray zone for categorization may result in the loss of important information about inhibitors.
4. Consideration of Chemical Family: The inclusion of the chemical family attribute significantly enhances the classification models. However, obtaining this attribute might require manual annotation or inspection as automatic extraction from databases like ChemBL might not be straightforward. Despite the effort required, incorporating this attribute contributes to the models' effectiveness.
5. State of Art in Urease Inhibitors: To the best of our knowledge, in the context of drug discovery targeting *Hp* infection through urease inhibition, our study stands out as the most comprehensive and systematic evaluation of optimal conditions for developing predictive models of bioactivity for potential inhibitor candidates. By

rigorously testing various attribute selection methods, machine learning algorithms, and bioactivity categorization strategies, we provide a robust framework that could significantly accelerate the identification and development of novel urease inhibitors. The elucidation of the structure–activity relationship (SAR) is crucial for rational drug design as it provides valuable information about how changes in the chemical structure of compounds affect their biological activity. Our investigation contributes to this understanding by identifying molecular features that correlate with UIs. By analyzing these relationships, researchers can gain insights into the chemical properties that are essentials for designing potent UIs.

6. Practical Significance for Drug Design: Our approach serves as a practical guide applicable not only to urease but also to other proteins in drug design, potentially impacting the field with its systematic methodology and comprehensive evaluation. This intersection between computational modeling and biological relevance highlights the significance of our findings in advancing both drug discovery efforts and our understanding of urease inhibition mechanisms. Developing predictive models that accurately classify compounds based on their inhibitory activity against any relevant clinical target, as demonstrated in our study, enables the efficient screening of large compound libraries to identify promising drug candidates. This can significantly accelerate the drug discovery process by prioritizing compounds with the highest likelihood of exhibiting inhibitory activity.

In general, this study illustrates the effectiveness of combining seven ML algorithms, three attribute selection methods, and six different strategies for categorizing inhibitory activity to enhance the prediction of urease inhibitors. The provided recommendations offer practical guidance for researchers aiming to develop effective classification models for similar biochemical systems.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25084303/s1.

Author Contributions: Conceptualization, D.B. and A.P.; methodology, N.M. and E.V.-M.; software, P.V.-H.; validation, C.G. and Á.C.-G.; formal analysis, D.B. and O.Y.; investigation, D.B., A.P. and N.M.; resources, J.M.P. and C.G.; data curation, J.G.; writing—original draft preparation, D.B. and A.P.; writing—review and editing, D.B. and A.P.; visualization, O.Y. and E.V.-M.; supervision, D.B. and A.P.; project administration, J.M.P.; funding acquisition, J.M.P., D.B. and E.V.-M. All authors have read and agreed to the published version of the manuscript.

Funding: D.B. would like to offer thanks to ANID FONDECYT de Iniciación #11220444 and FOVI230136. J.P-O thanks ANID FONDECYT de Iniciación #11201049. A.P. offers thanks to ANID FONDECYT Regular project No. 1220241. The APC was funded by FOVI230136 and Doctorado en Biotecnología Traslacional, Centro de Biotecnología de los Recursos Naturales, Universidad Católica del Maule, Talca, Chile.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Minkara, M.S.; Weaver, M.N.; Merz, K.M. Effect of 10.5 M Aqueous Urea on *Helicobacter pylori* Urease: A Molecular Dynamics Study. *Biochemistry* **2015**, *54*, 4121–4130. [CrossRef] [PubMed]
2. Kafarski, P.; Talma, M. Recent Advances in Design of New Urease Inhibitors: A Review. *J. Adv. Res.* **2018**, *13*, 101–112. [CrossRef] [PubMed]
3. Liu, Q.; Meng, X.; Li, Y.; Zhao, C.N.; Tang, G.Y.; Li, S.; Gan, R.Y.; Li, H. Bin Natural Products for the Prevention and Management of *Helicobacter pylori* Infection. *Compr. Rev. Food Sci. Food Saf.* **2018**, *17*, 937–952. [CrossRef] [PubMed]
4. Bravo, D.; Hoare, A.; Soto, C.; Valenzuela, M.A.; Quest, A.F. *Helicobacter pylori* in Human Health and Disease: Mechanisms for Local Gastric and Systemic Effects. *World J. Gastroenterol.* **2018**, *24*, 3071–3089. [CrossRef] [PubMed]

5. González, I.; Araya, P.; Roj, A. *Helicobacter pylori* Infection and Lung Cancer: New Insights and Future Challenges. *Chin. J. Lung Cancer* **2018**, *21*, 658–662. [CrossRef]
6. Hooi, J.K.Y.; Lai, W.Y.; Ng, W.K.; Suen, M.M.Y.; Underwood, F.E.; Tanyingoh, D.; Malfertheiner, P.; Graham, D.Y.; Wong, V.W.S.; Wu, J.C.Y.; et al. Global Prevalence of *Helicobacter pylori* Infection: Systematic Review and Meta-Analysis. *Gastroenterology* **2017**, *153*, 420–429. [CrossRef]
7. Vaira, D.; Holton, J.; Miglioli, M.; Menegatti, M.; Mule, P.; Barbara, L. Peptic Ulcer Disease and *Helicobacter pylori* Infection. *Curr. Opin. Gastroenterol.* **1994**, *10*, 98–104. [CrossRef]
8. Kavitt, R.T.; Lipowska, A.M.; Anyane-Yeboa, A.; Gralnek, I.M. Diagnosis and Treatment of Peptic Ulcer Disease. *Am. J. Med.* **2019**, *132*, 447–456. [CrossRef]
9. Wijarnpreecha, K.; Thongprayoon, C.; Nissaisorakarn, P.; Jaruvongvanich, V.; Nakkala, K.; Rajapakse, R.; Cheungpasitporn, W. Association of *Helicobacter pylori* with Chronic Kidney Diseases: A Meta-Analysis. *Dig. Dis. Sci.* **2017**, *62*, 2045–2052. [CrossRef]
10. Kim, T.H.; Cheung, D.Y. *Helicobacter pylori* Eradication, a Gordian Knot for Idiopathic Thrombocytopenic Purpura? *Gut Liver* **2016**, *10*, 323–324. [CrossRef]
11. Kim, B.J.; Kim, H.S.; Jang, H.J.; Kim, J.H. *Helicobacter pylori* Eradication in Idiopathic Thrombocytopenic Purpura: A Meta-Analysis of Randomized Trials. *Gastroenterol. Res. Pract.* **2018**, *2018*, 6090878. [CrossRef] [PubMed]
12. Kato, S.; Osaki, T.; Kamiya, S.; Zhang, X.S.; Blaser, M.J. *Helicobacter pylori* SabA Gene Is Associated with Iron Deficiency Anemia in Childhood and Adolescence. *PLoS ONE* **2017**, *12*, e0184046. [CrossRef]
13. Doorakkers, E.; Lagergren, J.; Engstrand, L.; Brusselaers, N. *Helicobacter pylori* Eradication Treatment and the Risk of Gastric Adenocarcinoma in a Western Population. *Gut* **2018**, *67*, 2092–2096. [CrossRef] [PubMed]
14. Shapla, U.M.; Raihan, J.; Islam, A.; Alam, F.; Solayman, N.; Gan, S.H.; Hossen, S.; Khalil, I. Propolis: The Future Therapy against *Helicobacter pylori*-Mediated Gastrointestinal Diseases. *J. Appl. Biomed.* **2018**, *16*, 81–99. [CrossRef]
15. Floch, P.; Mégraud, F.; Lehours, P. *Helicobacter pylori* Strains and Gastric MALT Lymphoma. *Toxins* **2017**, *9*, 132. [CrossRef]
16. Violeta Filip, P.; Cuciureanu, D.; Sorina Diaconu, L.; Maria Vladareanu, A.; Silvia Pop, C. MALT Lymphoma: Epidemiology, Clinical Diagnosis and Treatment. *J. Med. Life* **2018**, *11*, 187–193. [CrossRef]
17. Salar, A. Gastric MALT Lymphoma and *Helicobacter pylori*. *Med. Clin.* **2019**, *152*, 65–71. [CrossRef]
18. Flores-Treviño, S.; Mendoza-Olazarán, S.; Bocanegra-Ibarias, P.; Maldonado-Garza, H.J.; Garza-González, E. *Helicobacter pylori* Drug Resistance: Therapy Changes and Challenges. *Expert Rev. Gastroenterol. Hepatol.* **2018**, *12*, 819–827. [CrossRef] [PubMed]
19. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Parkin, D.M.; Piñeros, M.; Znaor, A.; Bray, F. Cancer Statistics for the Year 2020: An Overview. *Int. J. Cancer* **2021**, *149*, 778–789. [CrossRef]
20. Malfertheiner, P.; Megraud, F.; O'Morain, C.; Gisbert, J.P.; Kuipers, E.J.; Axon, A.; Bazzoli, F.; Gasbarrini, A.; Atherton, J.; Graham, D.Y.; et al. Management of *Helicobacter pylori* Infection-the Maastricht V/Florence Consensus Report. *Gut* **2017**, *66*, 6–30. [CrossRef]
21. Liou, J.M.; Fang, Y.J.; Chen, C.C.; Bair, M.J.; Chang, C.Y.; Lee, Y.C.; Chen, M.J.; Chen, C.C.; Tseng, C.H.; Hsu, Y.C.; et al. Concomitant, Bismuth Quadruple, and 14-Day Triple Therapy in the First-Line Treatment of *Helicobacter pylori*: A Multicentre, Open-Label, Randomised Trial. *Lancet* **2016**, *388*, 2355–2365. [CrossRef] [PubMed]
22. Hsieh, M.T.; Chang, W.L.; Wu, C.T.; Yang, H.B.; Kuo, H.Y.; Lin, M.Y.; Cheng, H.C.; Tsai, Y.C.; Sheu, B.S. Optimizing the MIC Breakpoints of Amoxicillin and Tetracycline for Antibiotic Selection in the Rescue Therapy of H. Pylori with Bismuth Quadruple Regimen. *Eur. J. Clin. Pharmacol.* **2020**, *76*, 1581–1589. [CrossRef] [PubMed]
23. Horie, R.; Handa, O.; Ando, T.; Ose, T.; Murakami, T.; Suzuki, N.; Sendo, R.; Imamoto, E.; Itoh, Y. *Helicobacter pylori* Eradication Therapy Outcome according to Clarithromycin Susceptibility Testing in Japan. *Helicobacter* **2020**, *25*, e12698. [CrossRef] [PubMed]
24. Alomari, M.; Taha, M.; Imran, S.; Jamil, W.; Selvaraj, M.; Uddin, N.; Rahim, F. Design, Synthesis, in Vitro Evaluation, Molecular Docking and ADME Properties Studies of Hybrid Bis-Coumarin with Thiadiazole as a New Inhibitor of Urease. *Bioorg. Chem.* **2019**, *92*, 103235. [CrossRef]
25. Kazmi, M.; Khan, I.; Khan, A.; Halim, S.A.; Saeed, A.; Mehsud, S.; Al-Harrasi, A.; Ibrar, A. Developing New Hybrid Scaffold for Urease Inhibition Based on Carbazole-Chalcone Conjugates: Synthesis, Assessment of Therapeutic Potential and Computational Docking Analysis. *Bioorg. Med. Chem.* **2019**, *27*, 115123. [CrossRef] [PubMed]
26. Modolo, L.V.; da-Silva, C.J.; Brandão, D.S.; Chaves, I.S. A Minireview on What We Have Learned about Urease Inhibitors of Agricultural Interest since Mid-2000s. *J. Adv. Res.* **2018**, *13*, 29–37. [CrossRef]
27. Liu, H.; Wang, Y.; Lv, M.; Luo, Y.; Liu, B.M.; Huang, Y.; Wang, M.; Wang, J. Flavonoid Analogues as Urease Inhibitors: Synthesis, Biological Evaluation, Molecular Docking Studies and In-Silico ADME Evaluation. *Bioorg. Chem.* **2020**, *105*, 104370. [CrossRef] [PubMed]
28. Chandra Babu, T.M.; Rajesh, S.S.; Bhaskar, B.V.; Devi, S.; Rammohan, A.; Sivaraman, T.; Rajendra, W. Molecular Docking, Molecular Dynamics Simulation, Biological Evaluation and 2D QSAR Analysis of Flavonoids from *Syzygium alternifolium* as Potent Anti-*Helicobacter pylori* Agents. *RSC Adv.* **2017**, *7*, 18277–18292. [CrossRef]
29. Yener, I.; Kocakaya, S.O.; Ertas, A.; Erhan, B.; Kaplaner, E.; Oral, E.V.; Yilmaz-Ozden, T.; Yilmaz, M.A.; Ozturk, M.; Kolak, U. Selective In Vitro and In Silico Enzymes Inhibitory Activities of Phenolic Acids and Flavonoids of Food Plants: Relations with Oxidative Stress. *Food Chem.* **2020**, *327*, 127045. [CrossRef]
30. Taha, M.; Rahim, F.; Khan, A.A.; Anouar, E.H.; Ahmed, N.; Shah, S.A.A.; Ibrahim, M.; Zakari, Z.A. Synthesis of Diindolylmethane (DIM) Bearing Thiadiazole Derivatives as a Potent Urease Inhibitor. *Sci. Rep.* **2020**, *10*, 7969. [CrossRef]

31. Khan, I.; Khan, A.; Ahsan Halim, S.; Saeed, A.; Mehsud, S.; Csuk, R.; Al-Harrasi, A.; Ibrar, A. Exploring Biological Efficacy of Coumarin Clubbed Thiazolo[3,2–b][1,2,4]triazoles as Efficient Inhibitors of Urease: A Biochemical and In Silico Approach. *Int. J. Biol. Macromol.* **2020**, *142*, 345–354. [CrossRef] [PubMed]
32. Sokmen, B.B.; Gumrukcuoglu, N.; Ugras, S.; Ugras, H.I.; Yanardag, R. Synthesis, Antibacterial, Antielastase, Antiurease and Antioxidant Activities of New Methoxy Substitued Bis-1,2,4-Triazole Derivatives. *J. Enzym. Inhib. Med. Chem.* **2013**, *28*, 72–77. [CrossRef]
33. Khan, I.; Ali, S.; Hameed, S.; Rama, N.H.; Hussain, M.T.; Wadood, A.; Uddin, R.; Ul-Haq, Z.; Khan, A.; Ali, S.; et al. Synthesis, Antioxidant Activities and Urease Inhibition of Some New 1,2,4-Triazole and 1,3,4-Thiadiazole Derivatives. *Eur. J. Med. Chem.* **2010**, *45*, 5200–5207. [CrossRef] [PubMed]
34. Moghimi, S.; Goli-Garmroodi, F.; Allahyari-Devin, M.; Pilali, H.; Hassanzadeh, M.; Mahernia, S.; Mahdavi, M.; Firoozpour, L.; Amanlou, M.; Foroumadi, A. Synthesis, Evaluation, and Molecular Docking Studies of Aryl Urea-Triazole-Based Derivatives as Anti-Urease Agents. *Arch. Pharm.* **2018**, *351*, 2–9. [CrossRef]
35. Salehi Ashani, R.; Azizian, H.; Sadeghi Alavijeh, N.; Fathi Vavsari, V.; Mahernia, S.; Sheysi, N.; Biglar, M.; Amanlou, M.; Balalaie, S. Synthesis, Biological Evaluation and Molecular Docking of Deferasirox and Substituted 1,2,4-Triazole Derivatives as Novel Potent Urease Inhibitors: Proposing Repositioning Candidate. *Chem. Biodivers.* **2020**, *17*, e1900710. [CrossRef] [PubMed]
36. Menteşe, E.; Akyüz, G.; Emirik, M.; Baltaş, N. Synthesis, in Vitro Urease Inhibition and Molecular Docking Studies of Some Novel Quinazolin-4(3H)-One Derivatives Containing Triazole, Thiadiazole and Thiosemicarbazide Functionalities. *Bioorg. Chem.* **2019**, *83*, 289–296. [CrossRef] [PubMed]
37. Salar, U.; Nizamani, A.; Arshad, F.; Khan, K.M.; Fakhri, M.I.; Perveen, S.; Ahmed, N.; Choudhary, M.I. Bis-Coumarins; Non-Cytotoxic Selective Urease Inhibitors and Antiglycation Agents. *Bioorg. Chem.* **2019**, *91*, 103170. [CrossRef]
38. Rashid, U.; Rahim, F.; Taha, M.; Arshad, M.; Ullah, H.; Mahmood, T.; Ali, M. Synthesis of 2-Acylated and Sulfonated 4-Hydroxycoumarins: In Vitro Urease Inhibition and Molecular Docking Studies. *Bioorg. Chem.* **2016**, *66*, 111–116. [CrossRef]
39. Naz, F.; Kanwal; Latif, M.; Salar, U.; Khan, K.M.; Al-Rashida, M.; Ali, I.; Ali, B.; Taha, M.; Perveen, S. 4-Oxycoumarinyl Linked Acetohydrazide Schiff Bases as Potent Urease Inhibitors. *Bioorg. Chem.* **2020**, *105*, 104365. [CrossRef]
40. Khan, K.M.; Iqbal, S.; Lodhi, M.A.; Maharvi, G.M.; Zia-Ullah; Choudhary, M.I.; Atta-ur-Rahman; Perveen, S. Biscoumarin: New Class of Urease Inhibitors; Economical Synthesis and Activity. *Bioorg. Med. Chem.* **2004**, *12*, 1963–1968. [CrossRef]
41. Faisal, M.; Saeed, A.; Shahzad, D.; Fattah, T.A.; Lal, B.; Channar, P.A.; Mahar, J.; Saeed, S.; Mahesar, P.A.; Larik, F.A. Enzyme Inhibitory Activities an Insight into the Structure–Activity Relationship of Biscoumarin Derivatives. *Eur. J. Med. Chem.* **2017**, *141*, 386–403. [CrossRef] [PubMed]
42. Wouters, O.J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018. *JAMA* **2020**, *323*, 844–853. [CrossRef]
43. Tang, Y.; Zhu, W.; Chen, K.; Jiang, H. New Technologies in Computer-Aided Drug Design: Toward Target Identification and New Chemical Entity Discovery. *Drug Discov. Today Technol.* **2006**, *3*, 307–313. [CrossRef] [PubMed]
44. Gertrudes, J.C.; Maltarollo, V.G.; Silva, R.A.; Oliveira, P.R.; Honorio, K.M.; da Silva, A.B.F. Machine Learning Techniques and Drug Design. *Curr. Med. Chem.* **2012**, *19*, 4289–4297. [CrossRef] [PubMed]
45. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1538. [CrossRef]
46. Trofymchuk, O.S.; Ortega, D.E.; Cortés-Arriagada, D.; Pereira, A.; Daniliuc, C.G.; Klitzke, C.F.; Santos, L.S.; Rojas, R.S. Neutral and Cationic Methallyl Nickel Complexes in Alkene Activation: A Combined DFT, ESI-MS and Chemometric Approach. *Catal. Sci. Technol.* **2021**, *11*, 7475–7485. [CrossRef]
47. Pereira, A.; Albornoz, C.; Trofymchuk, O.S. Data-Driven Analysis of Reactions Catalyzed by [CoCp*(CO)I2]. *Organometallics* **2022**, *41*, 1158–1166. [CrossRef]
48. Pereira, A.; Trofymchuk, O.S. Machine Learning Prediction of High-Yield Cobalt- and Nickel-Catalyzed Borylations. *J. Phys. Chem. C* **2023**, *127*, 12983–12994. [CrossRef]
49. Patel, L.; Shukla, T.; Huang, X.; Ussery, D.W.; Wang, S. Machine Learning Methods in Drug Discovery. *Molecules* **2020**, *25*, 5277. [CrossRef]
50. Adeshina, Y.O.; Deeds, E.J.; Karanicolas, J. Machine Learning Classification Can Reduce False Positives in Structure-Based Virtual Screening. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 18477–18488. [CrossRef]
51. Liu, H.X.; Zhang, R.S.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR and Classification Models of a Novel Series of COX-2 Selective Inhibitors: 1,5-Diarylimidazoles Based on Support Vector Machines. *J. Comput. Aided Mol. Des.* **2004**, *18*, 389–399. [CrossRef] [PubMed]
52. Roy, R.K.; Patra, N. Prediction of COMT Inhibitors Using Machine Learning and Molecular Dynamics Methods. *J. Phys. Chem. B* **2021**, *126*, 3477–3492. [CrossRef]
53. Tinivella, A.; Pinzi, L.; Rastelli, G. Prediction of Activity and Selectivity Profiles of Human Carbonic Anhydrase Inhibitors Using Machine Learning Classification Models. *J. Cheminform.* **2021**, *13*, 18. [CrossRef]
54. Mermer, A. Design, Synthesize and Antiurease Activity of Novel Thiazole Derivatives: Machine Learning, Molecular Docking and Biological Investigation. *J. Mol. Struct.* **2020**, *1222*, 128860. [CrossRef]
55. Aniceto, N.; Albuquerque, T.S.; Bonifácio, V.D.B.; Guedes, R.C.; Martinho, N. Using Machine Learning and Molecular Docking to Leverage Urease Inhibition Data for Virtual Screening. *Int. J. Mol. Sci.* **2023**, *24*, 8180. [CrossRef] [PubMed]

56. *Schrödinger Release 2021-1*; Maestro, Schrödinger, LCC: New York, NY, USA, 2024.
57. Shelley, J.C.; Cholleti, A.; Frye, L.L.; Greenwood, J.R.; Timlin, M.R.; Uchimaya, M. Epik: A Software Program for PKa Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691. [CrossRef] [PubMed]
58. Guha, R. Chemical Informatics Functionality in R. *J. Stat. Softw.* **2007**, *18*, 1–16. [CrossRef]
59. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria; Available online: https://www.R-project.org (accessed on 7 April 2024).
60. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
61. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
62. Medina-Franco, J.L.; Chávez-Hernández, A.L.; López-López, E.; Saldívar-González, F.I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41*, e2200116. [CrossRef]
63. Cihan Sorkun, M.; Mullaj, D.; Vianney, A.; Koelman, J.M.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization. *ChemRxiv* **2022**. [CrossRef]
64. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
65. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA; pp. 785–794.
66. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]