microorganisms

MDPI

Article

Harnessing Machine Learning to Uncover Hidden Patterns in Azole-Resistant CYP51/ERG11 Proteins

Otávio Guilherme Gonçalves de Almeida and Marcia Regina von Zeska Kress *

Faculdade de Ciências Farmacêuticas de Ribeirao Preto, Universidade de São Paulo, Ribeirão Preto 14040-903, SP, Brazil; otavio.almeida@usp.br
* Correspondence: kress@fcfrp.usp.br

Abstract: Fungal resistance is a public health concern due to the limited availability of antifungal resources and the complexities associated with treating persistent fungal infections. Azoles are thus far the primary line of defense against fungi. Specifically, azoles inhibit the conversion of lanosterol to ergosterol, producing defective sterols and impairing fluidity in fungal plasmatic membranes. Studies on azole resistance have emphasized specific point mutations in CYP51/ERG11 proteins linked to resistance. Although very insightful, the traditional approach to studying azole resistance is time-consuming and prone to errors during meticulous alignment evaluation. It relies on a reference-based method using a specific protein sequence obtained from a wild-type (WT) phenotype. Therefore, this study introduces a machine learning (ML)-based approach utilizing molecular descriptors representing the physiochemical attributes of CYP51/ERG11 protein isoforms. This approach aims to unravel hidden patterns associated with azole resistance. The results highlight that descriptors related to amino acid composition and their combination of hydrophobicity and hydrophilicity effectively explain the slight differences between the resistant non-wild-type (NWT) and WT (nonresistant) protein sequences. This study underscores the potential of ML to unravel nuanced patterns in CYP51/ERG11 sequences, providing valuable molecular signatures that could inform future endeavors in drug development and computational screening of resistant and nonresistant fungal lineages.

Keywords: CYP51; ERG11; machine learning; azoles; fungal resistance

Citation: Almeida, O.G.G.d.; von Zeska Kress, M.R. Harnessing Machine Learning to Uncover Hidden Patterns in Azole-Resistant CYP51/ERG11 Proteins. *Microorganisms* **2024**, *12*, 1525. https://doi.org/10.3390/microorganisms12081525

Academic Editors: Petros Ioannou and Diamantis P. Kofteridis

Received: 10 July 2024
Revised: 21 July 2024
Accepted: 22 July 2024
Published: 25 July 2024

1. Introduction

Fungal resistance to antifungals is a public health concern because limited resources are available for the treatment of mycoses [1]. In light of the increasing risk of fungal infections and the growing challenges associated with resistance and treatability, the World Health Organization has issued the inaugural list of priority fungal pathogens. This list serves as a guide for research, development, and public health initiatives, aiming to enhance knowledge acquisition and foster global comprehension and response to fungal-related concerns, among other critical objectives [2]. Much is known about the mechanisms of resistance to azoles, which is considered a multifactorial process [1,3] related not only to a single gene but also to efflux pumps, metabolic activity of the fungus, differential ergosterol compositionality, and mechanisms related to stress response, such as chaperone gene expression [3]. Comprehending this intricate multifactorial process proves challenging in laboratory settings. Many experiments involving the deletion of crucial genes involved in metabolic pathways are required to assess their role in resistance and the integration of multiple cross-related mechanisms from a dynamic perspective. Therefore, the study of single-gene-coding effectors recognized as a part of the resistance-mediated process remains persistent in the literature [1,4].

Among the extensively studied effectors of azole resistance are the CYP51 and ERG11 genes in filamentous fungi and yeast, respectively. These are homologous genes and are interchangeably treated as synonymous genes. Azoles act on membrane integrity by interfering with

14α-demethylases, also known as CYP51p or ERG11p, which are members of the cytochrome P450 monooxygenase superfamily [5,6]. These enzymes catalyze a crucial step in the biosynthesis of ergosterol. The inhibition of ergosterol biosynthesis is known to result in the accumulation of ergosterol precursors in the cell membrane, causing alterations in its fluidity and permeability and ultimately affecting the overall membrane structure. Additionally, the accumulation of deleterious sterols impairs membrane-bound enzymes, including chitin-synthase, and enzymes involved in detoxifying reactive oxygen species [5,6].

Several studies have focused on comparing CYP51 and ERG11 gene and protein sequences through global alignments to unravel patterns of conserved regions and their potential relevance to azole resistance. The study by [7] evaluated and furnished a comprehensive review evaluating the presence of amino acid substitutions in CYP51 sequences of *Aspergillus fumigatus* and the association between those substitutions and the resistance susceptibility profile. Similar approaches have been described for *Candida albicans* [8], Mucormycetes [9], the *Fusarium solani* species complex [10], and various other fungi [11,12].

Recent advancements in artificial intelligence (AI), particularly in machine learning (ML), have significantly enhanced the performance of data-driven applications across diverse domains. The field of deep neural networks and deep learning has played a crucial role in this progress. Recently, there has been a growing prevalence of ML models and advanced deep learning methods in the health sciences domain [13]. This increasing significance can be attributed to the remarkable progress in ML and the development of data-driven products. The availability of extensive structured and unstructured data, especially clinical and experimental data, has further fueled this trend [14]. In particular, the health field has experienced substantial benefits from adopting AI-driven solutions [15]. These advancements have been instrumental in various aspects of clinical decision-making and the management of infectious diseases [16]. Although there have been promising outcomes in hospital settings [17,18], antibiotic prescribing and management are exceptions. Beyond traditional stewardship programs, the importance of ML and deep learning has risen notably in addressing the antimicrobial resistance (AMR) challenge [19]. There is a call for continued investigation in this field to take advantage of recent advancements in ML and deep learning for a more robust approach to tackling the issue of AMR. Today, ML is a powerful tool for identifying hidden patterns in complex datasets and is now a reality in microbiology [20].

The random forest algorithm is a supervised learning algorithm [21] that requires training data to discern patterns and determine parameters. These parameters, numerical values derived from the data, constitute the foundation of a mathematical model. The optimization process involves varying numeric values, such as those related to random forest depth and sample splitting, through a grid search. This ensures the identification of optimal hyperparameters, improving model performance. Regarding supervised learning, splitting the dataset into training and test data is convenient. Typically, the training data comprise 20% to 30% of the original dataset, while the test data constitute 70% to 80%. The latter is used for model evaluation using dedicated metrics (e.g., accuracy, precision, and recall). To prepare numerical data from biological sequences such as DNA, RNA, or proteins, analysts must process these sequences to obtain significant numerical values reflecting various biological properties of the molecules [22].

A series of molecular descriptors can numerically represent protein sequences, most of which refer to physiochemical properties such as hydrophobicity, the composition of amino acids, and their relative frequencies in one-, di-, and tri-amino acids, or polarity, isoelectric point, probability of substitution using BLOSUM matrices, and others. This numerical description facilitates the comparison of diverse characteristics of biological sequences, extending beyond their structural aspects [23]. The resulting data can be summarized into a feature matrix suitable for supervised learning purposes.

ML methods employ algorithms to learn and predict AMR phenotypes directly from patients' clinical, demographic, and living-condition data [19]. Thus, this approach has been extended to sequenced bacterial genomes [24] and adapted to MALDI-TOF MS data,

enabling the detection of antifungal resistance in species such as *C. albicans* and *Aspergillus flavus* [25,26], among other applications.

In the context of fungal CYP51/ERG11 protein sequences, which were characterized by a notable degree of conservation among fungi, employing ML to identify novel attributes in these sequences is valuable. This approach can help address inquiries such as "What physiochemical properties in CYP51/ERG11 proteins might partially account for azole resistance in fungal strains?". These questions pose a challenge when relying solely on visual inspection of multiple global alignments of protein sequences. In addition, the hotspot-based approach in ML, which involves the presence of determined amino acids in conserved regions of alignments, lacks the mathematical background to assess differences among these conserved sequences quantitatively. The differentiation among these proteins can be quantified using protein descriptors, which are configured to mathematically describe and evaluate the molecular characteristics of proteins. These descriptors encompass numerical values summarizing various molecular properties, including charge, molecular weight, isoelectric point, polarity, hydrophobicity, frequency of amino acids, or values derived from algorithmic techniques reconstructing n-dimensional structures (i.e., 2-D or 3-D structures) of proteins [27]. Thus, descriptors have become valuable tools for effectively representing the molecular characteristics of CYP51 and ERG11 protein sequences in a manner conducive to the use of ML algorithms for extracting patterns and learning from the data. This approach goes beyond properties related to the global alignment of proteins for hotspot identification or the calculation of identity and similarity among closely related sequences.

This study aimed to unravel hidden patterns related to physiochemical descriptors of the CYP51/ERG11 proteins utilizing supervised learning ML algorithms. These patterns could provide insights into often overlooked resistance-related properties within protein sequences obtained from susceptible (wild-type phenotypes) and resistant (non-wild-type phenotypes) fungal strains. The information obtained from the data aims to enhance our comprehension of the functional disparities between a protein derived from a WT strain and one derived from an NWT strain. Additionally, the model developed in this study could be employed to characterize CYP51/ERG11 proteins and their isoforms, particularly for screening purposes.

2. Materials and Methods

2.1. CYP51 and ERG11 Amino Acid Sequences

Protein sequences of fungal CYP51 and ERG11 were selected from the National Center for Biotechnology (NCBI) protein database (Protein [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004—[cited 1 August 2023]. Available from: https://www.ncbi.nlm.nih.gov/protein/). The search focused on publications describing CYP51/ERG11 DNA sequences deposited on NCBI and associated MICs. The criteria for selection were: (i) sequences derived from Sanger DNA sequencing and documented in published papers, and (ii) provides minimum inhibitory concentration (MIC) data linked to the originating sequence isolates, excluding *Fusarium* spp. (also referred to as *Neocosmospora* spp. by some authors) [28]. The accession numbers for the predicted proteins were used to batch-download the protein sequences using the Entrez direct command line tool. In total, 282 protein sequences with associated MIC values were successfully obtained.

In addition, all MIC data were revisited and re-evaluated to ensure adherence to standards and achieve precise sequence categorization into wild-type (WT) and non-wild-type (NWT) epidemiological cutoff values (ECVs) criteria. This process followed the standard reference values described in the CLSI and EUCAST protocols to avoid misestimations in subsequent analyses. The ECVs established by Espinel-Ingroff et al. (2016) [29] were used for *Fusarium* spp. (*Neocosmospora* spp.) as the reference values for classifying protein sequences as WT or NWT. For *Candida* spp. and *Aspergillus* spp., the ECV was according to protocol M59, CLSI [30]. Additionally, two publications [31,32] containing *Aspergillus fumigatus* whole-genome sequences were selected. These studies recorded NWT and WT genomes

based on single-nucleotide polymorphism and MIC surveys. Subsequently, raw reads from 236 genomes (see Supplementary Table S1) were downloaded from the BioProject database of NCBI (https://www.ncbi.nlm.nih.gov/bioproject/; accessed on 1 March 2023) using accession numbers referenced in the selected publications. The reads underwent quality filtering with the bbduk tool (http://sourceforge.net/projects/bbmap/, accessed on 15 March 2023) using the parameters "hdist=1 tpe tbo qtrim=rl trimq=30 maq=30 minlen=90" and were assembled using Spades v3.13.1 [33] with the parameter "--careful". A custom database of CYP51, including its isoforms and ERG11 protein sequences, was built using DIAMOND v2.0.7 [34], utilizing the 282 protein sequences retrieved previously. Subsequently, gene and protein prediction for the assembled genomes was performed using the Funannotate tool v1.8.15 with parameters "--species "Aspergillus"", and "--busco_seed_species aspergillus_fumigatus", which incorporate the tools Augustus 3.5.0 [35] and Genemark [36] for ab initio and supervised learning predictions, respectively. The predicted proteins were then subjected to a DIAMOND [34] BLASTP algorithm against the custom database, with the best matches (evalue $\leq$ 0.001 and maximum identity) filtered for CYP51 isoform identification in the predicted proteins of the genomes. Finally, CYP51 sequences were obtained using the seqtk tool (https://github.com/lh3/seqtk; accessed on 1 March 2023) with the parameter "subseq" to retrieve the respective proteins annotated via blasting from the FASTA files. The resulting proteins (n = 472) and their metadata were combined with the publicly available proteins retrieved from NCBI into a single FASTA file to construct the feature table for ML purposes. The experimental design is summarized in Figure 1.

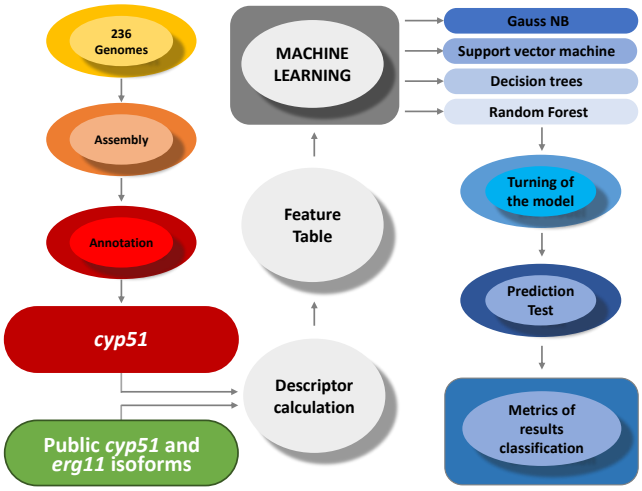

Figure 1. Experimental design for CYP51/ERG11 isoforms obtained from public databases and machine learning modeling.

In terms of representativeness, the dataset encompasses sequences from various fungal species: *Aspergillus awamori* (n = 11; 11 WT), *A. flavus* (n = 63; 36 WT and 27 NWT), *A. fumigatus* (n = 538; 211 WT and 327 NWT), *A. lucuhensis* (n = 2; 2 WT), *A. niger* (n = 11; 10 WT and 1 NWT), *A. tubingensis* (n = 12; 12 WT), *Candida glabrata* (n = 23; 8 WT and 15 NWT), *C. krusei* (n = 4; 1 WT and 3 NWT), *C. parapsilopsis* (n = 18; 1 WT and 17 NWT), *C. tropicalis* (n = 51; 22 WT and 29 NWT), *Fusarium keratoplasticum* (n = 15; 15 NWT), *Neocosmospora falciformis* (n = 2; 2 NWT), and *N. suttoniana* (n = 4; 4 NWT). The accession numbers for the CYP51 and ERG11 genes along with their cognate proteins and associated metadata are summarized in Supplementary Table S1.

2.2. Sequences Processing and Descriptors Calculations

The protein sequences were evaluated in terms of quality to exclude those presenting unrecognizable amino acids, such as the "X" character, which corresponds to ambiguous base calls ("N") in sequenced DNA bases. To achieve this goal, a custom script, check_aa.py, was used to identify and exclude sequences with "X" characters, result-

ing in a refined FASTA file suitable for subsequent analyses. After processing spurious sequences, 340 WT and 409 NWT protein sequences were selected for further analysis (Supplementary Table S1).

The proteins' descriptors were calculated using the R package 2.30.1 protr [37] and a Python 3.10.10 package written originally in R named peptides [38]. Various descriptors were calculated using the protr R package, including amino acid composition (AAC), dipeptide composition (DC), tripeptide composition (TC), composition (CTDC), transition (CTDT), and distribution (CTDD) of encoded classes in the protein amino acid sequences, as well as pseudo-amino acid composition (PAAC) and amphiphilic pseudo-amino acid composition (APAAC). Additionally, using the peptide python 3.10.10 package, several quantitative structure–activity relationship (QSAR) descriptors were calculated, such as BLOSUM indices, Cruciani properties, FASGAI vectors, Kidera factors, MS-WHIM scores, PCP descriptors, ProtFP descriptors, Sneath vectors, ST-scales, T-scales, VHSE-scales, and Z-scales.

2.3. Feature Table Building and Supervised Machine Learning Analysis

Supervised ML was employed in this study to classify CYP51 and ERG11 protein sequences into WT and NWT groups based on chemical signatures within the sequences. The feature table was constructed by performing a total joint of two tables generated from processing protein sequences using selected tools for descriptors calculations, all within a Google Colab notebook. For the ML analysis, the Scikit-learn library [39] and its methods were used in the Colab environment.

The following steps were taken to build a model employing four supervised ML algorithms (random forest, support vector machines (SVM), decision trees, and GaussNB). First, the feature table was split into a training set (30% of the data) and a test set (70% of the data). Then, various data normalization methods (StandardScaler, MinMaxScaler, RobustScaler, QuantileTransformer with the options "normal" and "uniform", and Normalizer) were evaluated for each ML algorithm on the training dataset to determine the most effective scaling method. Following the selection of the best scaling method for each algorithm based on their performance to achieve the highest accuracy, four learning models (one for each ML algorithm) were built and evaluated using five rounds of cross-validation in terms of accuracy, using the receiver operating characteristics and area under the curve (ROC–AUC) method and Matthews correlation coefficient (MCC score). Additionally, the algorithms were also compared in terms of accuracy on both the training and test datasets, and the algorithm exhibiting the best performance in both sets was selected for further model optimization.

2.4. Model Optimization and Evaluation

The optimal model, exhibiting the best performance, was generated using the random forest algorithm, as detailed in the upcoming Section 3. To further optimize this model, a grid search with cross-validation (GridSearchCV method) was performed by varying the following hyperparameters: "max_depth" (None, 2, 4, 8, 10, 12), "min_samples_split" (2, 5, 10), "criterion" (gini, entropy, log_loss), "max_features" (sqrt, log2, None), and "bootstrap" (False, True). The chosen scaling method for this optimization process was StandardScaler. After optimization, the evaluation of the model was performed using a classification report matrix. This matrix, displaying metrics such as accuracy, precision, recall, and f1 score, was plotted using the Seaborn library in Python. The metrics were derived by comparing the predictions between the training and test datasets. Moreover, a confusion matrix was plotted using the matplotlib library in Python, facilitating a comparison of false-positive, false-negative, true-positive, and true-negative predictions based on the labels WT and NWT.

2.5. Feature Importance Identification and Permutation Importance Analysis

Feature importance was determined via the random forest algorithm during the learning process. The top 20 most important features of the learning process are summarized in

a bar plot report, highlighting their mean decrease in impurity (MDI) metric. To compute the weight of each feature in terms of accuracy for both the training and test datasets, a permutation importance analysis was conducted through five repetitions utilizing the permutation_importance method from the Sklearn (v1.0) Python (v3.10.10) package for machine learning. However, it is worth noting that permutation analysis tends to overestimate the importance of continuous or high-cardinality categorical variables [40]. To address this issue, the entire dataset was processed before the permutation analysis. Initially, Spearman correlation was performed to detect highly correlated features (>95%) for removal, reducing the dataset dimensionality. This processing reduced the original feature table of 8772 attributes to 6184 features. Subsequently, the random forest model, utilizing the previously optimized parameters, was constructed to compute the importance of the feature. Boxplots were plotted to show the variation ranges of each feature's importance using the matplotlib library.

2.6. Global Alignment and Logo Construction

Given the homologous nature and evolutionary conservation of the CYP51 and ERG11 protein sequences [6,41,42], a comprehensive alignment of all 749 protein sequences was built using Clustal Omega v1.2.4 [43]. Subsequently, a logo showing conserved and variable regions among these sequences was generated using the WebLogo online tool v3.7.12 [44].

3. Results

The feature table consisted of 749 proteins derived from Sanger DNA sequencing of several fungal species exhibiting distinct susceptibility profiles to azoles (Supplementary Table S1). The DNA sequences available on NCBI were utilized with the Entrez direct tool to retrieve predicted and annotated CYP51 and ERG11 protein sequences based on the GenBank accession numbers of DNA sequences. In the case of *A. fumigatus* whole-genome sequences, a prediction step was necessary to identify CYP51 homologs from the 282 proteins directly obtained from NCBI. This approach yielded 340 WT and 409 NWT protein sequences characterized by their physiochemical properties using dedicated descriptors. The final feature table comprised 8772 columns, each representing a descriptor's property (i.e., amino acid composition and hydrophobicity), and served as the foundation for constructing the ML model.

3.1. Machine Learning Model Construction

The creation of the ML model involved choosing the optimal supervised algorithm and scaling method. Figure 2A illustrates the relationships between the scaling method and ML algorithms. Specifically, for GaussianNB, all the scalers exhibited nearly equal performance, with the MinMax scaler slightly outperforming the others. Both random forest and decision trees demonstrated similar effectiveness across all scalers, and the Standard scaler was selected for both algorithms. The Uniform scaler emerged as the most effective option for support vector machines (SVM) (Figure 2A–C).

After scaling the training datasets with the appropriate scalers for each ML algorithm, the performance of each algorithm was assessed using metrics such as accuracy, ROC–AUC, and MCC (Figure 2B). The accuracy was evaluated as the ratio of correct predictions to total predictions (CP/TP). The ROC–AUC score, which reflects the area under the receiver operating characteristic curve, provides insights into model specificity and sensitivity, with the following interpretations: 0.5–0.6 (failed), 0.6–0.7 (worthless), 0.7–0.8 (poor), 0.8–0.9 (good), and >0.9 (excellent) [45]. Finally, the MCC score which is mathematically expressed as $MCC = (TP \times TN - FP \times FN)/(\sqrt{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)})$, in which TP = true positives, TN = true negatives, FN = false negatives, and FP = false positives, was considered. High MCC values are achieved when a model correctly predicts the majority of the metrics in a confusion matrix [45].

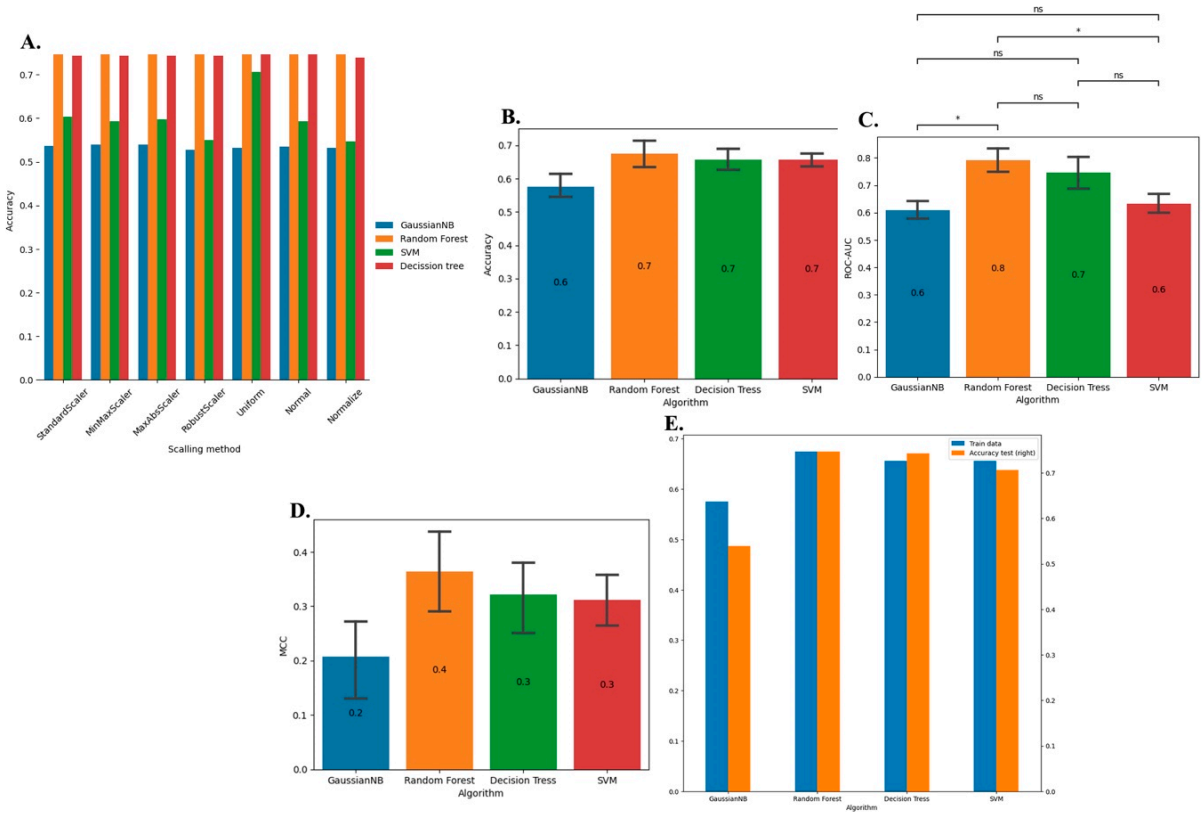

Figure 2. Modeling metrics. (**A**) Performance of scaling methods. (**B**) Accuracy of machine learning algorithms using Uniform, the best scaling method. (**C**) ROC–AUC of machine learning algorithms using Uniform, the best scaling method. *, *p*-value of ≤0.05; ns, *p*-value of >0.05 (Mann–Whitney multiple groups comparison with Bonferroni correction). (**D**) MCC of machine learning algorithms using Uniform, the best scaling method. (**E**) Comparison of accuracy metric between training and test datasets among machine learning algorithms.

Random forest showed more significant accuracy variation than did other algorithms, with decision trees and SVM exhibiting similar average accuracy values. The GaussianNB method demonstrated lower accuracy than the other methods (Figure 2B). Regarding ROC–AUC scores, the random forest model achieved the highest score at 80%, followed by the decision trees at 70%. The GaussianNB and SVM models exhibited comparable average ROC–AUC scores at 60% (Figure 2C). Despite the lower MCC scores for all algorithms, the random forest algorithm outperformed the others, with a score of 40%, surpassing the decision tree (30%), SVM (30%), and GaussianNB (20%) algorithms (Figure 2D). Notably, compared with the algorithms, random forests exhibited balanced accuracy for training and test datasets compared to the other algorithms, indicating a greater probability of generalized predictions (Figure 2E).

However, the statistical analysis, which was conducted through Mann–Whitney multiple groups comparison with Bonferroni correction, did not reveal any significant differences at a *p*-value of ≤0.05 among the algorithms concerning accuracy, ROC–AUC, or MCC. This lack of significance is evident from the deviations observed in the bar plots (Figure 2B). Consequently, any of these algorithms can be employed for model training, expecting similar results. However, we decided to use the random forest due to its slight superiority in terms of accuracy variation, higher ROC–AUC, better MCC scores, and consistent accuracy across both the training and test datasets (Figure 2B–E). Concerning model optimization, the random forest model underwent training based on five cross-validation rounds, resulting in the identification of the best parameters: "bootstrap": false, "criterion": entropy, "max_depth": 2, "max_features": log2, "min_samples_split": 2, achieving a max accuracy score of approximately 72%.

3.2. Model Evaluation

A confusion matrix is usually used to visualize the correct classification of labels. This study's labels refer to the WT and NWT phenotypes of CYP51 and ERG11 protein sequences associated with azole resistance. The confusion matrix is presented in Figure 3A. As illustrated, among the 287 NWT protein sequences, the model incorrectly assigned seven sequences as having the WT phenotype. Conversely, out of 238 WT sequences, 134 were wrongly assigned as NWT phenotypes. These findings suggest that the model is more inclined to correctly identify NWT phenotypes than WT phenotypes (Figure 3A).

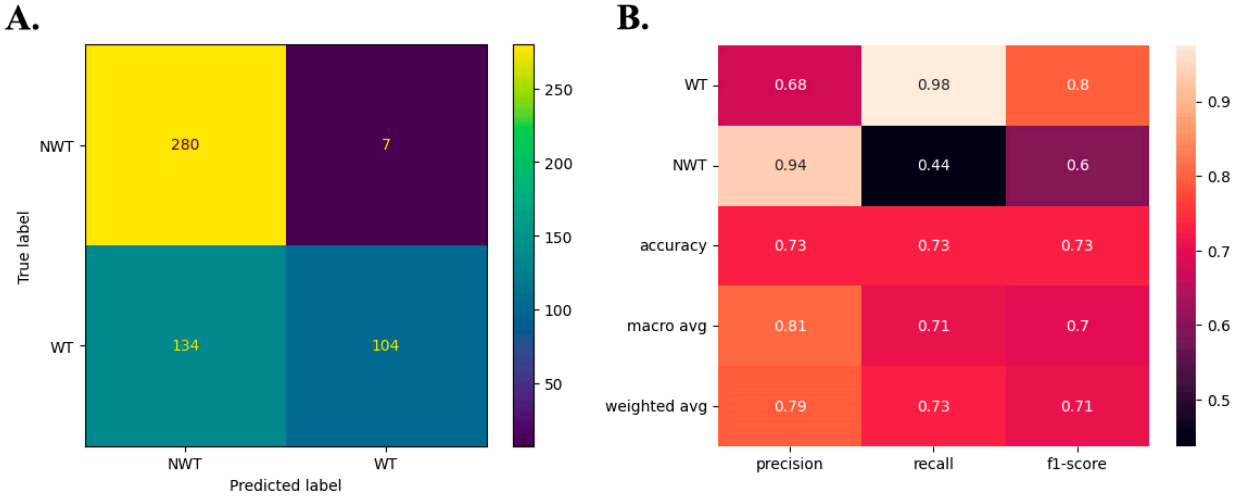

Figure 3. Metrics of supervised learning. (**A**) The confusion matrix shows correct and incorrect predictions for WT and NWT phenotypes, and (**B**) heatmap resuming the main classification metrics.

Figure 3B shows the classification results, illustrating key metrics for evaluating ML: accuracy, precision, recall, and f1-score. Precision, used to measure the proportion of truly positive predictions among those predicted as positive (defined as TP/(TP + FP), where TP and FP denote true positives and false positives, respectively), indicates that the model accurately predicted 94% of the NWT phenotypes and only 68% of the WT phenotypes (Figure 3B).

Recall, defined as TP/(TP + FN), where FN represents false negatives, aims to disclose the proportion of true positives not correctly predicted via the model. As depicted in Figure 3B, NWT's (44%) recall is lower than WT's (98%) recall. This suggests that, while the prediction of NWT sequences was highly accurate, the model may have overlooked many NWT sequences. These scores indicate that the model can correctly identify NWT sequences in a superior way than it can identify WT sequences. However, the model may have failed to identify some NWT sequences, contributing to an increased number of false-WT (false-negative) phenotypes. The f1-score, calculated as the harmonic mean of precision and recall (2*(Precision*Recall)/(Precision + Recall)), serves to harmonize measurements of precision and recall by assigning equal weight to both metrics, offering a comprehensive view of predictions. As depicted in Figure 3B, the f1-score for WT phenotype prediction was high at 80%. This suggests that although the model predicts the true phenotype for WT sequences with moderate precision, it can inaccurately predict some NWT sequences as WT sequences, as corroborated by the data in Figure 3A (indicating a high number of false-negative or false-WT phenotypes).

In contrast, the low recall for NWT prediction implies that the model may occasionally fail to identify all true NWT phenotypes in the dataset despite exhibiting high precision. Although the model can accurately identify NWT phenotypes, it might misclassify sequences with ambiguous or low-specific signatures as WT. Consequently, the precision in identifying WT sequences diminishes as the recall rate increases. This signifies that the reduction in precision primarily stems from an increase in false positives, specifically, false-WT sequences. The mean accuracy was 73%, indicating that most predictions were correct, considering the labeled test dataset.

Macro and weighted-average metrics represent each measurement's arithmetic and weighted means across both classes (WT and NWT), respectively. Each class is assigned a weight based on the number of predicted sequences in the weighted average. As shown in Figure 3B, both averages are close for all the assessed metrics. This suggests that, despite some shortcomings in predicting specific phenotypes, either due to a loss of precision or an increase in false negatives (for NWT phenotypes) or false positives (for WT phenotypes), the model consistently achieved an average accuracy above 70%. Therefore, it can be considered suitable for discovering patterns associated with azole resistance in CYP51 and ERG11 sequences. The model also applied to screening CYP51 and ERG11 sequences lacking associated MIC data for classification. Additionally, it can be used to compare novel protein sequences using in silico methods.

3.3. Descriptors for the Classification of CYP51 and ERG11 Sequences

The top 20 descriptors used for the classification of the CYP51 and ERG11 sequences are shown in Figure 4A. These descriptors, listed in descending order of relative importance, include physiochemical property descriptors based on multidimensional scaling (PCP) represented by the attribute E4 and pseudo-amino acid composition (PseAAC), with the attributes Xc1.T, Xc1.P, Xc2.lambda.15, and Xc2.lambda.29; structural topology scale (ST-scale) descriptor, represented by the attribute ST1; tripeptide composition descriptor, with the attributes VTA, IPA, LVA, VFE, ISY, TRW, LNG, VIF, and GVP; dipeptide composition descriptor, represented by the attribute YD; distribution descriptor, represented by prop.5.G1.residue.50; factor analysis descriptor, represented by the attribute F4; and amphiphilic pseudo-amino acid composition (APseAAC) descriptor, represented by the attribute Pc2.hydrophobicity.8.

Figure 4B illustrates the importance of the features in the training dataset, revealing variations in the tripeptide composition descriptors (KET, NPL, ALL, IKE, and EGE), distribution descriptors (prop1.G2.residue.50, prop7.G1.residue75, and prop.7.G3.residue25), APseAAC descriptors (Pc2.hydrophobicity.13, Pc2.hydrophobicity.14, Pc2.hydrophobicity.12, Pc2.hydrophobicity.16, Pc2.hydrophobicity.15, Pc1.K, and Pc1.N), PseAAC descriptors (Xc2.lambda.8, Xc2.lambda.6, Xc2.lambda.15, and Xc1.F), dipeptide descriptors (PI and LA), transition descriptors (prop7.Tr1221, prop1.Tr1331, and prop1.Tr2332), composition descriptor (polarizability.Group2), frequency of amino acid descriptor (attribute: C), Kidera factor descriptor (attribute: KF8), and ProtFP descriptor (attribute: ProtFP4).

Figure 4C presents the impact of the permutation shuffling of values in the test dataset. The main descriptors that varied included the distribution descriptor (attributes: prop1.G2.residue.50 and prop7.G1.residue.75), tripeptide frequency descriptor (attributes: ALL, NPL, and KET), PseAAC descriptor (attributes: Xc2.lambda.29, Xc1.F, and Xc2.lambda.8), dipeptide frequency descriptor (attributes: GM and NE), and APseAAC descriptor (attributes: Pc2.hydrophobicity.13 and Pc2.hydrophobicity.16).

Collectively, these descriptors and their attributes aim to address crucial questions: Does a chemical signature exist in these sequences that allows their characterization in terms of resistance? How might variations in their values impact the estimation? The first assignment was answered by elucidating the top 20 descriptor attributes in the dataset. The second assignment was approached through permutation analysis, where the values of all descriptor attributes were randomly varied, and the resultant effect on model accuracy was determined on both the training and test datasets (Figure 4B,C). Thus, combining the consensual importance of descriptor attributes, as inferred from the permutation analysis on both the training and test datasets, it becomes apparent that the variation induced by shuffling values in descriptors such as PseAAC, tripeptide composition, dipeptide composition, distribution descriptor, and APseAAC has the potential to describe hidden physiochemical properties in azole-resistant CYP51 and ERG11 proteins. These properties may explain minor differences between these sequences and shed light on their impact on fungal susceptibility to azoles.

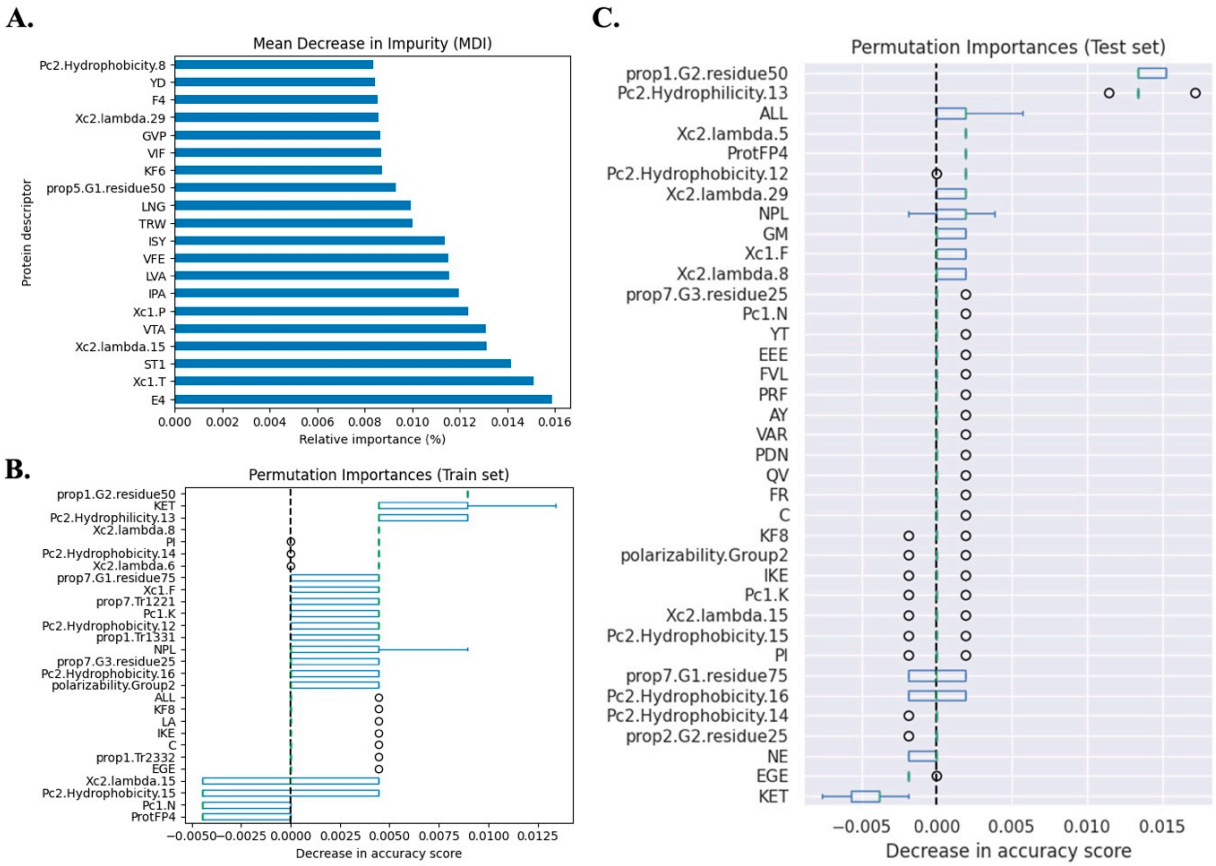

Figure 4. Most important (top 20) features' attributes. (**A**) Main attributes related to WT and NWT phenotypes used in random forest classifier. Permutation importance analysis on (**B**) train and (**C**) test datasets.

3.4. Conservation among Sequences

A global alignment was conducted on all 749 selected protein sequences to address bias toward identifying NWT sequences. This analysis aimed to unveil the reasons behind the model's evaluation metrics favoring NWT over WT. Then, a logo was created to visualize conserved motifs among the sequences and understand how such conservation might complicate the differentiation of closely related sequences associated with strains exhibiting different susceptibility levels to azole antifungals. Therefore, a biological interpretation of the model accuracy was facilitated by comparing global alignments, shedding light on the model's robust discriminatory power in classifying NWT phenotypes compared to its ability to discern WT sequences.

Figure 5 shows the conservation of amino acid positions in CYP51 and ERG11. The first five amino acids varied among the sequences, showing no discernible pattern. Similar variability is observed at positions 93–97, 455–463, and 559–562. Conversely, the remaining positions exhibit high conservation, with minimal variations in amino acid prevalence (e.g., position 40, position 186, and position 345) and complete conservation in specific sequences (e.g., position range of 160–164). The high degree of conservation among the sequences and sparse differences in amino acid composition at specific positions may explain the model's inability to predict WT sequences, as protein sequences tend to be highly conserved (Figure 5).

Figure 5. The protein logo shows the conservation degree of motifs shared by CYP51 and ERG11 amino acids' sequences.

4. Discussion

This study aimed to develop a machine learning (ML) model for identifying azole-resistant strains. The ML model discerns key features distinguishing between the susceptible (WT) and resistant (NWT) fungal strains, focusing on protein descriptors to gain insights into the molecular resistance mechanisms associated with CYP51/ERG11 protein sequences. Identifying these crucial features contributes to a deeper understanding of the biological factors influencing azole resistance, providing valuable insights for the scientific community.

The modeling process effectively mirrors real-world scenarios faced by researchers. The random forest algorithm was chosen for modeling because it demonstrated superior performance to GaussNB, decision trees, and support vector machines. Although there were no significant differences in accuracy among these algorithms, random forest demonstrated consistently superior results across multiple metrics used for model evaluation, including MCC and ROC–AUC analysis. This underscores its efficacy as the preferred option for predicting azole-resistant proteins.

In general, the optimized model presented satisfactory accuracy, aligning well with the challenges faced by wet-lab peers. This alignment is particularly crucial when dealing with conserved regions overlapping in CYP51/ERG11 protein sequences, as it may impair the differentiation of crucial attributes between the WT and NWT phenotypes. The achieved accuracy of 72% is deemed realistic, acknowledging that no model can correctly predict 100% of all the data. Conversely, an accuracy lower than 60% when the model is evaluated using testing data may reflect overfitting, a phenomenon caused by excessive learning from the training data, limiting its ability to generalize effectively to external datasets. Therefore, establishing literature guidelines on the minimal accuracy score for ML models is crucial because different fields may require distinct cutoff values. Analysts must exercise judgment in determining a suitable threshold for the model's accuracy, considering the context of data acquisition, balancing, and structure [46]. It is also imperative to consider alternative metrics, such as the precision, recall, and f1-score, as they contribute to determining the effectiveness of the proposed model.

In this study, the ML model better predicted NWT sequences than WT sequences, primarily due to the high precision scores exhibited in NWT predictions. This observation was supported with the confusion matrix analysis, revealing a superior discriminatory ability in classifying NWT sequences. However, the recall for NWT predictions was lower than that for WT predictions, indicating that the model was somewhat restrictive in identifying specific NWT sequences incorrectly assigned as WT sequences. This increases the number of false-positive WT sequences, contributing to a decrease in precision for WT sequence predictions. Nevertheless, when comparing both the recall and precision scores summarized with the f1-score, the model demonstrated a tendency to discriminate between the WT and NWT phenotypes effectively in most instances. Furthermore, these test data may explain the comparatively lower performance of the model as determined via the MCC score on the training data (MCC score of 40%). The decrease in precision (true positives) is linked to this decrease in the MCC score despite the model presenting satisfactory (good) ROC–AUC and accuracy scores on the training datasets.

Despite using varied numerical values to describe the 749 proteins and constructing a substantial feature matrix with 8772 columns (attributes), discerning patterns proved challenging due to the high degree of conservation among the sequences. Nevertheless, even with the slight variation in amino acid composition, as illustrated by the protein logos, the model effectively classified many WT and NWT proteins. The metric scores used for model evaluation are deemed realistic.

It is important to acknowledge that the CYP51 and ERG11 sequences used in the present study originated from different fungal species, each of which presented distinct profiles of azole susceptibility. This implies that a particular isolate or strain may be resistant to voriconazole but susceptible to itraconazole and posaconazole or vice versa (Supplementary Table S1). Consequently, as several protein sequences presented different

levels of susceptibility to different azole classes, this may account for the greater number of false-positive WT phenotypes, resulting in low precision for WT classification. In essence, the overlap of highly conserved sequences associated with different azole resistance profiles is a confounding factor. Additionally, the determination of MIC is subjective and visually assessed [4], introducing potential bias into the analysis. In certain instances, the MIC may be fungistatic rather than fungicidal [47]. Furthermore, the dataset comprised various publications with different references, such as CLSI and EUCAST, with the ECV values as the basis for the compilation. This approach might have led to inaccuracies in the dataset concerning NWT or WT strain classification. Despite diligent re-evaluation by comparing MICs to the reference guideline threshold, the analysis may have been influenced by the analyst's subjectivity, potentially introducing inherent bias.

Fungal resistance is a complex process involving multiple factors [48] and cannot be underestimated. More than simply attributing a resistant or non-resistant phenotype to an entire microorganism based solely on the presence of a unique CYP51/ERG11 protein may be needed to understand the intricate mechanisms of resistance. Although distinctive substitution patterns in the amino acids of CYP51/ERG11 proteins are positively correlated with azole resistance [6], correlation alone does not necessarily indicate causality. Instead, it indicates the direction and strength of association, influenced by several factors [49], such as multifactorial resistance.

In the context of antifungal resistance, specific amino acid substitutions associated with resistant phenotypes may reflect a selection process favoring isoforms capable of adopting distinct conformational folds. This adaptation allows them to synergize with other mechanisms or metabolic pathways, effectively overcoming the stress induced by azole interference. Moving beyond this perspective, a feature importance analysis was used to determine whether shuffling a particular attribute leads to a decrease in an ML model's accuracy (or another scoring metric). This analysis played a crucial role in identifying features offering insights into biological patterns related to azole resistance to various fungal proteins. This study evaluated the feature importance of both training and test data. In the initial scenario, the aim was to visualize the most crucial features for model learning. Conversely, in the second scenario, the emphasis shifted to visualizing the most significant features for model prediction. Combining the most crucial features from both scenarios made it feasible to estimate the optimal features capable of distinguishing between WT and NWT sequences.

As observed, the PseAAC, tripeptide composition, dipeptide composition, distribution descriptor, factor analysis descriptor, and APseAAC descriptor emerged as the most important features impacting model predictions. PseAAC is a method that considers the composition of the 20 amino acids in a protein sequence and incorporates the order in which they appear, utilizing a combination of discrete sequence correlation factors and the traditional amino acid composition [50]. Similarly, APseAAC combines sequence correlation factors to distinguish hydrophobic and hydrophilic distribution patterns in a protein sequence [50].

Composition descriptors, here summarized by the composition of amino acids (n = 20), dipeptides (n = 400), and tripeptides (n = 8000), calculated as the percentage of a given amino acid or combinations of amino acids (di- and tripeptides) concerning the entire sequence, serve as valuable parameters for condensing the information into a single value. This allows for comparing sequences with varying lengths and facilitates pattern extraction for ML analysis [51]. Furthermore, distribution descriptors are used to compute the percentage of neutral, polar, and hydrophobic residues along the length of a protein sequence [52].

Some studies investigating the 3D structure of the CYP51/ERG11 protein have revealed differences in the conformations of the WT and NWT proteins. Specifically, these variations involve the interaction mode of certain azoles' lateral long-chains with 14-α-demethylase, affecting the channel that communicates substrates to the active sites of the protein. These active sites function as the loci for the interaction between azoles and

the enzyme [10,53–55]. Consequently, studies on CYP51/ERG11 contribute significantly to understanding the mechanisms of azole resistance in fungi, reflecting a synergistic multifactorial process. Within this context, differences in the order and frequency of specific amino acids, along with their combinations in the sequences of CYP51/ERG11 WT and NWT phenotypes, may lead to tertiary conformation changes associated with patterns of hydrophobicity and hydrophilic regions. These alterations modify the active site of CYP51/ERG11. Docking analysis has revealed that CYP51 active site consists of hydrophobic amino acid residues that interact with the hydrophobic lanosterol (the precursor of ergosterol) through π-π and π-alkylation contacts with the amino acid residues [56]. Additionally, maintaining a balance between hydrophobic and hydrophilic characteristics is necessary to ensure that the substrate and active site interact in the CYP51/ERG11 channel. As demonstrated by previous studies, an increase in hydrophilicity reduces the affinity of these interactions [57,58].

5. Conclusions

Recognizing the limitations of traditional approaches, ML has emerged as a powerful tool for extracting valuable insights from complex and intricate biological datasets. The mathematical modeling of CYP51/ERG11 sequences to construct feature tables is crucial for revealing hidden physiochemical patterns within these sequences. In this study, we focused on protein attributes related to amino acid composition and their combination and hydrophobicity and hydrophilicity. This analysis revealed slight differences between NWT and WT proteins, highlighting significant molecular signatures. These findings have promising implications for future drug development strategies or in silico screening of potential NWT and WT lineages through comprehensive whole-genome analysis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/microorganisms12081525/s1, Table S1: Metadata of datasets and respective publications selected in this study.

Author Contributions: O.G.G.d.A., conception and design of the work; the acquisition, analysis, and interpretation of data; drafted the work and substantively revised it; approved the submitted version. M.R.v.Z.K., conception; interpretation of data; substantively revised the manuscript; approved the submitted version. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The São Paulo Research Foundation (FAPESP), Grants #2022/00754-4, #2023/12463-7, and #2020/07546-2. The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) with funding code 310425/2021-2.

Data Availability Statement: The codes used in this study are available at https://github.com/Otavio20/CYPER accessed on 1 October 2023 along with the raw feature table of computed proteins' descriptors.

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

References

1. Arastehfar, A.; Gabaldón, T.; Garcia-Rubio, R.; Jenks, J.D.; Hoenigl, M.; Salzer, H.J.F.; Ilkit, M.; Lass-Flörl, C.; Perlin, D.S. Drug-Resistant Fungi: An Emerging Challenge Threatening Our Limited Antifungal Armamentarium. *Antibiotics* **2020**, *9*, 877. [CrossRef]
2. WHO. *WHO Fungal Priority Pathogens List to Guide Research, Development and Public Health Action*; WHO: Geneva, Switzerland, 2022.
3. Srinivasan, A.; Lopez-Ribot, J.L.; Ramasubramanian, A.K. Overcoming Antifungal Resistance. *Drug Discov. Today Technol.* **2014**, *11*, 65–71. [CrossRef] [PubMed]
4. Cuenca-Estrella, M. Antifungal Drug Resistance Mechanisms in Pathogenic Fungi: From Bench to Bedside. *Clin. Microbiol. Infect.* **2014**, *20*, 54–59. [CrossRef]
5. Becher, R.; Wirsel, S.G.R. Fungal Cytochrome P450 Sterol 14α-Demethylase (CYP51) and Azole Resistance in Plant and Human Pathogens. *Appl. Microbiol. Biotechnol.* **2012**, *95*, 825–840. [CrossRef]
6. Song, J.; Zhang, S.; Lu, L. Fungal Cytochrome P450 Protein Cyp51: What We Can Learn from Its Evolution, Regulons and *Cyp51*-Based Azole Resistance. *Fungal Biol. Rev.* **2018**, *32*, 131–142. [CrossRef]

7. Dudakova, A.; Spiess, B.; Tangwattanachuleeporn, M.; Sasse, C.; Buchheidt, D.; Weig, M.; Groß, U.; Bader, O. Molecular Tools for the Detection and Deduction of Azole Antifungal Drug Resistance Phenotypes in *Aspergillus* Species. *Clin. Microbiol. Rev.* **2017**, *30*, 1065–1091. [CrossRef]
8. Warrilow, A.G.; Nishimoto, A.T.; Parker, J.E.; Price, C.L.; Flowers, S.A.; Kelly, D.E.; Rogers, P.D.; Kelly, S.L. The Evolution of Azole Resistance in *Candida Albicans* Sterol 14α-Demethylase (CYP51) through Incremental Amino Acid Substitutions. *Antimicrob. Agents Chemother.* **2019**, *63*. [CrossRef]
9. Caramalho, R.; Tyndall, J.D.A.; Monk, B.C.; Larentis, T.; Lass-Flörl, C.; Lackner, M. Intrinsic Short-Tailed Azole Resistance in Mucormycetes Is Due to an Evolutionary Conserved Aminoacid Substitution of the Lanosterol 14α-Demethylase. *Sci. Rep.* **2017**, *7*, 3–12. [CrossRef] [PubMed]
10. Vermeulen, P.; Gruez, A.; Babin, A.L.; Frippiat, J.P.; Machouart, M.; Debourgogne, A. CYP51 Mutations in the Fusarium Solani Species Complex: First Clue to Understand the Low Susceptibility to Azoles of the Genus *Fusarium*. *J. Fungi* **2022**, *8*, 533. [CrossRef]
11. Sionov, E.; Chang, Y.C.; Garraffo, H.M.; Dolan, M.A.; Ghannoum, M.A.; Kwon-Chung, K.J. Identification of a *Cryptococcus Neoformans* Cytochrome P450 Lanosterol 14α-Demethylase (Erg11) Residue Critical for Differential Susceptibility between Fluconazole/Voriconazole and Itraconazole/Posaconazole. *Antimicrob. Agents Chemother.* **2012**, *56*, 1162–1169. [CrossRef]
12. Zhao, H.; Tao, X.; Song, W.; Xu, H.; Li, M.; Cai, Y.; Wang, J.; Duan, Y.; Zhou, M. Mechanism of *Fusarium Graminearum* Resistance to Ergosterol Biosynthesis Inhibitors: G443S Substitution of the Drug Target FgCYP51A. *J. Agric. Food Chem.* **2022**, *70*, 1788–1798. [CrossRef]
13. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
14. Anahtar, M.N.; Yang, J.H.; Kanjilal, S. Applications of Machine Learning to the Problem of Antimicrobial Resistance: An Emerging Model for Translational Research. *J. Clin. Microbiol.* **2021**, *59*. [CrossRef]
15. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A Guide to Deep Learning in Healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef]
16. Tran, N.K.; Albahra, S.; May, L.; Waldman, S.; Crabtree, S.; Bainbridge, S.; Rashidi, H. Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing. *Clin. Chem.* **2022**, *68*, 125–133. [CrossRef] [PubMed]
17. Oonsivilai, M.; Mo, Y.; Luangasanatip, N.; Lubell, Y.; Miliya, T.; Tan, P.; Loeuk, L.; Turner, P.; Cooper, B.S. Using Machine Learning to Guide Targeted and Locally-Tailored Empiric Antibiotic Prescribing in a Children's Hospital in Cambodia. *Wellcome Open Res.* **2018**, *3*, 131. [CrossRef]
18. Baysari, M.T.; Lehnbom, E.C.; Li, L.; Hargreaves, A.; Day, R.O.; Westbrook, J.I. The Effectiveness of Information Technology to Improve Antimicrobial Prescribing in Hospitals: A Systematic Review and Meta-Analysis. *Int. J. Med. Inform.* **2016**, *92*, 15–34. [CrossRef] [PubMed]
19. Elyan, E.; Hussain, A.; Sheikh, A.; Elmanama, A.A.; Vuttipittayamongkol, P.; Hijazi, K. Antimicrobial Resistance and Machine Learning: Challenges and Opportunities. *IEEE Access* **2022**, *10*, 31561–31577. [CrossRef]
20. Goodswen, S.J.; Barratt, J.L.N.; Kennedy, P.J.; Kaufer, A.; Calarco, L.; Ellis, J.T. Machine Learning and Applications in Microbiology. *FEMS Microbiol. Rev.* **2021**, *45*, fuab015. [CrossRef] [PubMed]
21. Singh, A.; Thakur, N.; Sharma, A. A Review of Supervised Machine Learning Algorithms. In Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 1310–1315.
22. Jiao, Y.; Du, P. Performance Measures in Evaluating Machine Learning Based Bioinformatics Predictors for Classifications. *Quant. Biol.* **2016**, *4*, 320–330. [CrossRef]
23. Fernández-Torras, A.; Comajuncosa-Creus, A.; Duran-Frigola, M.; Aloy, P. Connecting Chemistry and Biology through Molecular Descriptors. *Curr. Opin. Chem. Biol.* **2022**, *66*, 102090. [CrossRef]
24. Kim, J.I.; Maguire, F.; Tsang, K.K.; Gouliouris, T.; Peacock, S.J.; McAllister, T.A.; McArthur, A.G.; Beiko, R.G. Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clin. Microbiol. Rev.* **2022**, *35*, e00179-21. [CrossRef]
25. Delavy, M.; Cerutti, L.; Croxatto, A.; Prod'hom, G.; Sanglard, D.; Greub, G.; Coste, A.T. Machine Learning Approach for *Candida Albicans* Fluconazole Resistance Detection Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. *Front. Microbiol.* **2019**, *10*, 3000. [CrossRef]
26. Normand, A.-C.; Chaline, A.; Mohammad, N.; Godmer, A.; Acherar, A.; Huguenin, A.; Ranque, S.; Tannier, X.; Piarroux, R. Identification of a Clonal Population of Aspergillus Flavus by MALDI-TOF Mass Spectrometry Using Deep Learning. *Sci. Rep.* **2022**, *12*, 1575. [CrossRef] [PubMed]
27. Emonts, J.; Buyel, J.F. An Overview of Descriptors to Capture Protein Properties—Tools and Perspectives in the Context of QSAR Modeling. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 3234–3247. [CrossRef] [PubMed]
28. O'Donnell, K.; Al-Hatmi, A.M.S.; Aoki, T.; Brankovics, B.; Cano-Lira, J.F.; Coleman, J.J.; de Hoog, G.S.; Di Pietro, A.; Frandsen, R.J.N.; Geiser, D.M.; et al. No to *Neocosmospora*: Phylogenomic and Practical Reasons for Continued Inclusion of the *Fusarium solani* Species Complex in the Genus Fusarium. *mSphere* **2020**, *5*. [CrossRef]
29. Espinel-Ingroff, A.; Colombo, A.L.; Cordoba, S.; Dufresne, P.J.; Fuller, J.; Ghannoum, M.; Gonzalez, G.M.; Guarro, J.; Kidd, S.E.; Meis, J.F.; et al. International Evaluation of MIC Distributions and Epidemiological Cutoff Value (ECV) Definitions for *Fusarium* Species Identified by Molecular Methods for the CLSI Broth Microdilution Method. *Antimicrob. Agents Chemother.* **2016**, *60*, 1079–1084. [CrossRef]

30. *M59*; Epidemiological Cutoff Values for Antifungal Susceptibility Testing. CLSI: Wayne, PA, USA, 2018.
31. Rhodes, J.; Abdolrasouli, A.; Dunne, K.; Sewell, T.R.; Zhang, Y.; Ballard, E.; Brackin, A.P.; van Rhijn, N.; Chown, H.; Tsitsopoulou, A.; et al. Population Genomics Confirms Acquisition of Drug-Resistant *Aspergillus Fumigatus* Infection by Humans from the Environment. *Nat. Microbiol.* **2022**, *7*, 663–674. [CrossRef]
32. Abdolrasouli, A.; Rhodes, J.; Beale, M.A.; Hagen, F.; Rogers, T.R.; Chowdhary, A.; Meis, J.F.; Armstrong-James, D.; Fisher, M.C. Genomic Context of Azole Resistance Mutations in *Aspergillus Fumigatus* Determined Using Whole-Genome Sequencing. *mBio* **2015**, *6*. [CrossRef]
33. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
34. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef]
35. Stanke, M.; Morgenstern, B. AUGUSTUS: A Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints. *Nucleic Acids Res.* **2005**, *33*, W465–W467. [CrossRef]
36. Borodovsky, M.; Lomsadze, A. Eukaryotic Gene Prediction Using GeneMark.Hmm-E and GeneMark-ES. *Curr. Protoc. Bioinform.* **2011**, *35*, 4.6.1–4.6.10. [CrossRef] [PubMed]
37. Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. Protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [CrossRef]
38. Osorio, D.; Rondon-Villarreal, P.; Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J.* **2015**, *7*, 4–14. [CrossRef]
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
41. Zhang, J.; Li, L.; Lv, Q.; Yan, L.; Wang, Y.; Jiang, Y. The Fungal CYP51s: Their Functions, Structures, Related Drug Resistance, and Inhibitors. *Front. Microbiol.* **2019**, *10*, 691. [CrossRef]
42. Celia-Sanchez, B.N.; Mangum, B.; Brewer, M.; Momany, M. Analysis of Cyp51 Protein Sequences Shows 4 Major *Cyp51* Gene Family Groups across Fungi. *G3 Genes Genomes Genet.* **2022**, *12*, jkac249. [CrossRef]
43. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef]
44. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188–1190. [CrossRef]
45. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews Correlation Coefficient (Mcc) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Min.* **2021**, *14*, 1–22. [CrossRef] [PubMed]
46. Rácz, A.; Bajusz, D.; Héberger, K. Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules* **2019**, *24*, 2811. [CrossRef]
47. Lewis, J.S.; Graybill, J.R. Fungicidal vs. Fungistatic: What's in a Word? *Expert Opin. Pharmacother.* **2008**, *9*, 927–935. [CrossRef] [PubMed]
48. Roca, A.; Matilla, M.A. Microbial Antibiotics Take the Lead in the Fight against Plant Pathogens. *Microb. Biotechnol.* **2023**, *16*, 28–33. [CrossRef]
49. Roy-García, I.; Rivas-Ruiz, R.; Pérez-Rodríguez, M.; Palacios-Cruz, L. Correlation: Not all correlation entails causality. *Rev. Alerg. Mex.* **2019**, *66*, 354–360. [CrossRef]
50. Chou, K.-C. Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* **2005**, *21*, 10–19. [CrossRef]
51. Bhasin, M.; Raghava, G.P.S. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* **2004**, *279*, 23262–23266. [CrossRef]
52. Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8700–8704. [CrossRef]
53. Xiao, L.; Madison, V.; Chau, A.S.; Loebenberg, D.; Palermo, R.E.; McNicholas, P.M. Three-Dimensional Models of Wild-Type and Mutated Forms of Cytochrome P450 14α-Sterol Demethylases from *Aspergillus Fumigatus* and *Candida Albicans* Provide Insights into Posaconazole Binding. *Antimicrob. Agents Chemother.* **2004**, *48*, 568–574. [CrossRef] [PubMed]
54. Chunquan, S.; Zhenyuan, M.; Haitao, J.; Jianzhong, Y.; Wenya, W.; Xiaoying, C.; Guoqiang, D.; Jiaguo, L.; Wei, G.; Wannian, Z. Three-Dimensional Model of Lanosterol 14α-Demethylase from *Cryptococcus Neoformans*: Active-Site Characterization and Insights into Azole Binding. *Antimicrob. Agents Chemother.* **2009**, *53*, 3487–3495. [CrossRef]
55. Matowane, R.G.; Wieteska, L.; Bamal, H.D.; Kgosiemang, I.K.R.; Van Wyk, M.; Manume, N.A.; Abdalla, S.M.H.; Mashele, S.S.; Gront, D.; Syed, K. In Silico Analysis of Cytochrome P450 Monooxygenases in Chronic Granulomatous Infectious Fungus *Sporothrix Schenckii*: Special Focus on CYP51. *Biochim. Biophys. Acta Proteins Proteom.* **2018**, *1866*, 166–177. [CrossRef] [PubMed]

56. Sun, B.; Huang, W.; Liu, M.; Lei, K. Comparison and Analysis of the Structures and Binding Modes of Antifungal SE and CYP51 Inhibitors. *J. Mol. Graph. Model.* **2017**, *77*, 1–8. [CrossRef] [PubMed]
57. Schiaffella, F.; Macchiarulo, A.; Milanese, L.; Vecchiarelli, A.; Costantino, G.; Pietrella, D.; Fringuelli, R. Design, Synthesis, and Microbiological Evaluation of New *Candida albicans* CYP51 Inhibitors. *J. Med. Chem.* **2005**, *48*, 7658–7666. [CrossRef]
58. Verma, A.K.; Majid, A.; Hossain, M.S.; Ahmed, S.F.; Ashid, M.; Bhojiya, A.A.; Upadhyay, S.K.; Vishvakarma, N.K.; Alam, M. Identification of 1, 2, 4-Triazine and Its Derivatives Against Lanosterol 14-Demethylase (CYP51) Property of *Candida albicans*: Influence on the Development of New Antifungal Therapeutic Strategies. *Front. Med. Technol.* **2022**, *4*, 845322. [CrossRef]