Received: 15 May 2022 | Accepted: 16 August 2022

DOI: 10.1002/hep4.2088

ORIGINAL ARTICLE

Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3

Annika Behrendt[1] | **Pegah Golchin**[2] | **Filip König**[1] | **Daniel Mulnaes**[1] | **Amelie Stalke**[3,4] | **Carola Dröge**[5,6] | **Verena Keitel**[5,6] | **Holger Gohlke**[1,7]

[1]Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Darmstadt, Germany

[3]Department of Human Genetics, Hannover Medical School, Hannover, Germany

[4]Division of Kidney, Department of Pediatric Gastroenterology and Hepatology, Liver, and Metabolic Diseases, Hannover Medical School, Hannover, Germany

[5]Department for Gastroenterology, Hepatology, and Infectious Diseases, Medical Faculty, Otto von Guericke University, Magdeburg, Germany

[6]Department for Gastroenterology, Hepatology, and Infectious Diseases, University Hospital, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[7]John-von-Neumann-Institute for Computing, Jülich Supercomputing Center, Institute of Biological Information Processing (IBI-7: Structural Biochemistry), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Jülich, Germany

Correspondence
Holger Gohlke, Institute for Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany.
Email: gohlke@uni-duesseldorf.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01GM1904A and 01GM1904B

Abstract

The phosphatidylcholine floppase multidrug resistance protein 3 (MDR3) is an essential hepatobiliary transport protein. MDR3 dysfunction is associated with various liver diseases, ranging from severe progressive familial intrahepatic cholestasis to transient forms of intrahepatic cholestasis of pregnancy and familial gallstone disease. Single amino acid substitutions are often found as causative of dysfunction, but identifying the substitution effect in *in vitro* studies is time and cost intensive. We developed variant assessor of MDR3 (Vasor), a machine learning-based model to classify novel MDR3 missense variants into the categories benign or pathogenic. Vasor was trained on the largest data set to date that is specific for benign and pathogenic variants of MDR3 and uses general predictors, namely Evolutionary Models of Variant Effects (EVE), EVmutation, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, and PON-P2 along with other variant properties, such as half-sphere exposure and posttranslational modification site, as input. Vasor consistently outperformed the integrated general predictors and the external prediction tool MutPred2, leading to the current best prediction performance for MDR3 single-site missense variants (on an external test set: F1-score, 0.90; Matthew's correlation coefficient, 0.80). Furthermore, Vasor predictions cover the entire sequence space of MDR3. Vasor is accessible as a webserver at https://cpclab.uni-duesseldorf.de/mdr3_predictor/ for users to rapidly obtain prediction results and a visualization of the substitution site within the MDR3 structure. The MDR3-specific prediction tool Vasor can provide reliable predictions of single-site amino acid substitutions, giving users a fast way to initially assess whether a variant is benign or pathogenic.

INTRODUCTION

Bile formation is a carefully regulated system, from bile acid synthesis to secretion of bile acids across the canalicular membrane. Adenosine triphosphate (ATP)-binding cassette (ABC) transporters present on the canalicular membrane of hepatocytes are responsible for the transport of primary bile components, namely, bile acids through the bile salt export pump (BSEP, *ABCB11*), cholesterol through the ABC subfamily G members 5 and 8 (ABCG5/ABCG8), and phospholipids through multidrug resistance protein 3 (MDR3, *ABCB4*). MDR3 acts as a floppase, translocating substrates, such as phosphatidylcholine, from the inner to the outer membrane leaflet[1,2] and exposing the substrate for extraction into primary bile.[3] Recent studies have suggested different transport pathways that follow either an alternating two-site access model through the protein's inner cavity[4] or a credit-card swipe mechanism along transmembrane helix 7 (TM H7).[5] MDR3 dysfunction has been linked to various liver-associated diseases, including intrahepatic cholestasis of pregnancy, low phospholipid-associated cholelithiasis, drug-induced liver injury, progressive familial intrahepatic cholestasis type 3, liver fibrosis/cirrhosis, and hepatobiliary malignancy.[6–12]

It is estimated that at least 70% of disease-causing *ABCB4* variants are amino acid substitutions, whereas variants leading to premature stop codons and protein truncations are in the minority.[13] However, while the advancement of sequencing allows rapid testing of patients, it remains challenging for clinicians and researchers to assess the potential impact of novel missense variants.

Evaluation of newly found MDR3 amino acid substitutions by *in vitro* cellular assays remains time consuming. Machine-learning-based prediction tools instead offer rapid analysis and have led in recent years to many predictors.[14,15] Nonetheless, general predictors do not consistently perform well on all proteins, necessitating the development of protein-specific prediction tools. To date, there is no MDR3-specific predictor available for classifying amino acid substitutions despite the vital role of MDR3 in bile homeostasis. An initial evaluation of general predictor performances on MDR3 variants suggested MutPred as a well-performing tool[16,17]; however, generalization is difficult due to only 21 tested variants with established cellular effects. Additionally, the tested variants presented a clear bias toward pathogenic effects.

Here, we created an MDR3-specific variant data set and trained a machine-learning algorithm using established general prediction tools, namely Evolutionary Models of Variant Effects (EVE), EVmutation, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, and PON-P2,[18–24] as well as half-sphere exposure and posttranslational modification (PTM) site influence as features to obtain an MDR3-specific prediction tool for help in classifying variants as benign or pathogenic (see Figure 1 for a graphical overview). Our predictor, variant assessment of MDR3 (Vasor), performed better than each integrated general predictor. Additionally, Vasor outperformed MutPred2,[25] a general predictor we chose for comparison based on the suggested high performance of its predecessor MutPred on MDR3.[16] We provide easy access to Vasor through a webserver where users can enter a missense variant of interest and obtain a prediction if it is benign or pathogenic together with an estimate of the prediction probability. Additionally, the mutation site is displayed on the structure of MDR3, giving the user a comprehensive view of the local site and the overall position of the assessed variant.

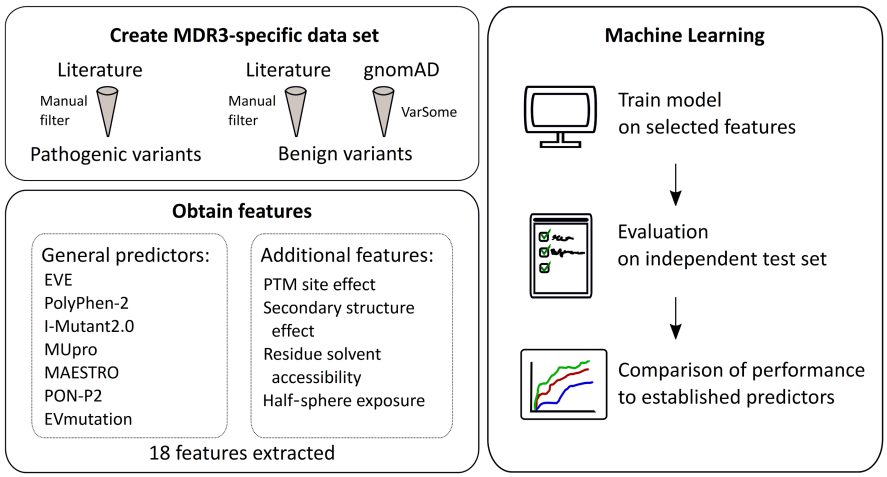

FIGURE 1 Graphical overview of data set generation and machine-learning approach. For details, see text. EVE, Evolutionary Models of Variant Effects; gnomAD, Genome Aggregation Database; MDR3, multidrug resistance protein 3; PTM, posttranslational modification.

MATERIALS AND METHODS

MDR3 missense variants

MDR3 variants were obtained from a literature search for variants causative of MDR3 dysfunction or known variants with no effect in any MDR3-associated disease (see Table S1). We excluded variants with unclear information on disease association (i.e., no *in vitro* verification analysis and no information on clinical indications for disease association) to eliminate false positives (FPs) or false negatives (FNs). As studied benign variants for MDR3 are rare,[13,16] further missense variants were obtained from Genome Aggregation Database (gnomAD) v2.1.1[26] to increase the number of benign variants. During the generation of the gnomAD database, individuals with severe pediatric diseases are removed; however, it is possible that pathogenic variants exist in the gnomAD data set. Accordingly, we employed a selection step to exclude FN cases of MDR3 variants. Using the platform VarSome,[27] variants were preclassified following the guidelines of The American College of Medical Genetics and Association for Molecular Pathology (ACMG-AMP)[28] rules, and variants with a likely pathogenic or pathogenic effect were removed, whereas variants with uncertain significance, likely benign, or benign classification by VarSome were integrated into the data set. These steps were included to create a high-quality data set to keep the number of misclassified variants low but at the same time retain a sufficiently high number of variants. The final list of variants contained 85 pathogenic and 279 benign variants. Every variant was mapped to the longest MDR3 isoform, corresponding to Uniprot[29] entry P21439-1.

Data set and features

The list of MDR3 variants was subjected to general predictors for missense mutations (EVE, PolyPhen-2, I-Mutant2.0, MUpro, MAESTRO, PON-P2, and EVmutation), and additional features (half-sphere exposure, secondary structure disruption, PTM site, and relative solvent accessibility) were computed, creating an MDR3-specific feature set.

EVE is a recently developed, unsupervised, computational method that trained Bayesian variational autoencoders on multiple sequence alignments to classify variant effects based on a computed evolutionary index followed by a fitted global–local mixture of Gaussian mixture models.[18] PolyPhen-2 employs a naive Bayes classifier for predicting variant effects using sequence-based features and structure-based features.[19] I-Mutant2.0 predicts protein stability changes by using a support vector machine-based tool trained on either sequence or structural information.[20] MUpro predicts stability changes on single-site mutations by using sequence and structural information with a support vector machine.[21] Both I-Mutant2.0 and MUpro predict the direction of stability change and the energy difference. MAESTRO employs a combination of machine-learning approaches to predict the energy difference introduced by missense mutations based on consensus, along with predicting a confidence score.[22] PON-P2 applies selected features from evolutionary conservation and biochemical properties of amino acids to develop a random forest classifier that classifies mutations as benign or pathogenic or those with unknown significance.[23] EVmutation explicitly considers interdependencies between residues or nucleotide bases in their unsupervised statistical method to include epistasis.[24]

EVE and EVmutation predictions for the MDR3 protein were accessed using the precomputed data set available from the method creators (https://evemodel.org/, https://marks.hms.harvard.edu/evmutation/human_proteins.html). I-Mutant2.0, Mupro, and MAESTRO predictions were generated using their standalone downloadable versions. PolyPhen-2 predictions were accessed using the batch query of the webserver (http://genetics.bwh.harvard.edu/pph2/bgi.shtml) with the default values. PON-P2 predictions were generated using the sequence submission feature for variants of the webserver (http://structure.bmc.lu.se/PON-P2/).

Additional features were added to explicitly integrate effects on PTM sites, variant location in α-helical or β-sheet secondary structure, and effects on residue solvent accessibility. Known PTM sites from the literature were supplemented by potential PTM sites predicted by PhosphoMotif,[30] PhosphoSitePlus,[31] NetPhos,[32] and the Eukaryotic Linear Motif (ELM) database.[33] The secondary structure was extracted from the MDR3 structure (Protein Data Bank identification [PDB ID]: 6S7P), using the database of secondary structure assignments DSSP.[34,35] Relative solvent accessibility was computed based on residue exposure calculated with DSSP divided by the maximal residue solvent accessibility.[36] Half-sphere exposure was introduced before[37] to measure residue solvent exposure and surpass limitations of relative solvent accessibility. It was implemented using values from the Biopython HSExposure module calculated according to the half-sphere corresponding to the direction of the sidechain of the residue as measured from the Cα atom.

Machine learning

The obtained data set was cleaned from non-numerical values. In the case of binary features, such as classification features of general predictors, −1 was set if no prediction was available to distinguish from benign (value 0) or pathogenic (value 1) predictions. Additionally, relative solvent accessibility and

half-sphere exposure were set to −1 if no prediction value was obtained in order to distinguish from prediction values of 0. Other numerical features were replaced by 0 if no prediction for the respective feature was available. The correlation between features within the data set was assessed by the Spearman *R* correlation coefficient.

A test set was generated by selecting 20 benign and 20 pathogenic variants from the overall data set. To avoid a bias toward specific amino acids, we minimized the root-mean-square deviation (RMSD)-based difference between the amino acid distribution of the variants within the test set compared to the overall data set (Figure S1). After randomly drawing 10 variants into the test set, the RMSD-based difference between the amino acid distribution of the general data set and current test set was computed; further variants were only transferred into the test set if they met one of the following conditions: (a) the RMSD between reference sequence and substituted amino acid distributions decreased by addition of the new variant, (b) the RMSD between reference sequence amino acid distributions decreased while the RMSD between substituted amino acid distributions did not increase more than 0.1, or (c) the RMSD between substituted amino acid distributions decreased while the RMSD between reference sequence amino acid distributions did not increase more than 0.1. Due to the limited size of the data set, it might not otherwise be possible to draw a variant for the test set. The test set was withheld from the machine-learning training step and used for final validation.

To handle the imbalance between the pathogenic (85 variants) and benign (279 variants) class, we used the synthetic minority oversampling technique (SMOTE).[38] This method generates new synthetic data points by using existing minority data points within the *N*-dimensional data set space, drawing lines to the five nearest minority class neighbors, and randomly selecting synthetic data points along these lines to balance out the classes.

On the training data set, the XGBoost algorithm[39] (as implemented in the Python library) was trained using the default gradient-boosted tree (gbtree); the maximum depth of a tree (max_depth) was 3, subsample 0.6, and step size (learning_rate) 0.02. The training was evaluated using repeated *k*-fold cross-validation, with $k = 3$ and the value of repeats (n_repeats) = 5. Using this procedure, the training data set was randomly split into three equally sized folds, where each fold is used as an internal test data set with the remaining two folds as training data sets. The performance results were measured and visualized in receiver operating characteristic (ROC) curves for comparison to the final test set. These steps were repeated 5 times.

To reduce features and estimate feature importance, we analyzed the tree-based feature importance and the permutation importance, leading to the removal of the four least informative features shared in both feature-importance measures: relative solvent accessibility, I-Mutant2.0 stability sign, I-Mutant2.0 deltaG value, and secondary structure disruption. Tree-based feature importance was computed using the XGBoost algorithm built-in feature and the "gain" (average gain across all splits where a feature is used). Permutation-based feature importance was computed by random shuffling each feature consecutively, followed by a performance test; this denoted performance alterations following feature permutation. The performance of the model without feature selection is shown in Figure S2.

The trained model, termed Vasor, predicts a probability ranging from 0 to 1 for a given variant to belong to the pathogenic class. Predictions above (below) 0.5 are classified as pathogenic (benign).

Comparison to established predictors

To assess the general performance of Vasor, we compared it to the general predictors EVE, PolyPhen-2, PON-P2, and MutPred2. MutPred2 predictions were used to compare our prediction tool to an external general predictor as MutPred2 was not used as an input feature for Vasor. The standalone version of MutPred2 was used to classify each variant within the entire data set, and a threshold of 0.5 was used to classify pathogenicity.[25] The performance of Vasor and the other predictors was evaluated on the entire data set and the test set. This ensured increased fairness for the performance comparison as Vasor may have an advantage over other predictors based on its training on the training data set. ROC and precision-recall curves were adjusted to the availability of variants each predictor was able to classify over the entire data set (i.e., if general predictors did not classify a variant into the category benign or pathogenic, the respective variant could not be assessed and curves were shown only on assessable variants). To account for this, the coverage of each predictor of the MDR3 data set was computed.

Performance evaluation

The performance of Vasor and the other prediction tools was evaluated using recommended measures for binary classifiers,[40] including additionally the F1-score as well as visualization in ROC and precision-recall curves. The measures are based on the values of correctly classified variants, indicated by true positives (TPs) for correctly predicted pathogenic variants and true negatives (TNs) for correctly predicted benign variants as well as incorrectly classified variants indicated by FPs for variants predicted as pathogenic

albeit benign and false negatives FNs for variants predicted as benign albeit pathogenic. The analyzed measures of recall, specificity, precision, negative predictive value (NPV), accuracy, F1-score, and Matthew's correlation coefficient (MCC) were calculated as

$$\text{Recall} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,TP}{2\,TP + FP + FN} \quad (6)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

Webserver tool

Vasor can be accessed online at https://cpclab.uni-duesseldorf.de/mdr3_predictor/. Users can enter a single-site amino acid missense MDR3 variant; the tool will only recognize MDR3 variants corresponding to the largest protein isoform UniProt ID: P21439-1. The entry needs to be in the format of the standard International Union of Pure and Applied Chemistry code for amino acids, entering first the one-letter code of the amino acid of the reference sequence, followed by the position and the amino acid substitution of interest. On the results page, users can see the predicted classification (either benign or pathogenic) and the probability of pathogenicity (PoP). This probability ranges from 0 (highest probability for the variant to be benign) to 1 (highest probability for the variant to be pathogenic). Probability values close to 0.5 indicate less confidence in the prediction.

Additionally, the results page displays the structure of the MDR3 protein (PDB ID: 6S7P) with the NGL Viewer,[41,42] including the membrane localization obtained from the Orientations of Proteins in Membranes database[43] as a red and blue plane. The substituted residue is colored according to the predicted effect either in red (pathogenic) or green (benign). The user can download a zip archive containing a high-resolution image of the complete protein, PDB files of the reference sequence and the variant protein, and high-resolution images of the position with the reference sequence residue or the substituted one.

Code availability

The code for Vasor was written in Python 3.9 and is provided for download at https://cpclab.uni-duesseldorf.de/index.php/Software.

RESULTS

Generation of a data set with informative features and good overall coverage of the MDR3 protein

To establish an MDR3-specific prediction tool, we prepared a data set of benign and pathogenic MDR3 variants. Relevant literature on MDR3-associated diseases was screened. Variants with unclear association to effects were omitted to avoid misclassified variants. Additionally, the gnomAD database[26] was screened for MDR3 variants, and the results were subjected to filtering by VarSome[27] using ACMG-AMP rules[28] to remove variants with a high potential for a pathogenic effect. This step was necessary as pathogenic MDR3 variants on a single allele with a potential late-onset or mild phenotype might have been included in the gnomAD database. Next, we used general predictors (EVE,[18] EVmutation,[24] PolyPhen-2,[19] I-Mutant2.0,[20] MUpro,[21] MAESTRO,[22] and PON-P2[23]) and descriptors of the variant site, namely, the disruption of secondary structure, possible PTM site disturbance, and changes in the relative solvent accessibility and half-sphere exposure of the position in question, as features in the data set. Projecting the variant locations from the data set onto the known cryogenic electron microscopy structure of MDR3 (PDB ID: 6S7P)[4] revealed a broad coverage of the structure with benign and pathogenic variants (Figure 2A). No functional domain is devoid of variants, and we do not observe large clusters of benign or pathogenic variants, which may indicate a potential bias within the data set. Such a bias might prevent applying the tool to areas of low coverage. Hence, we expect that our tool can generalize predictions to every position of MDR3.

To further probe for domains of low applicability, we mapped variants misclassified by Vasor to the MDR3 structure. Misclassified variants from the data set tend to occur on the solvent-exposed surface of the protein rather than within buried regions of the protein (Figure S3). As solvent-exposed residues are less evolutionary conserved than buried residues,[44] the obtained trend might visualize the underlying increased

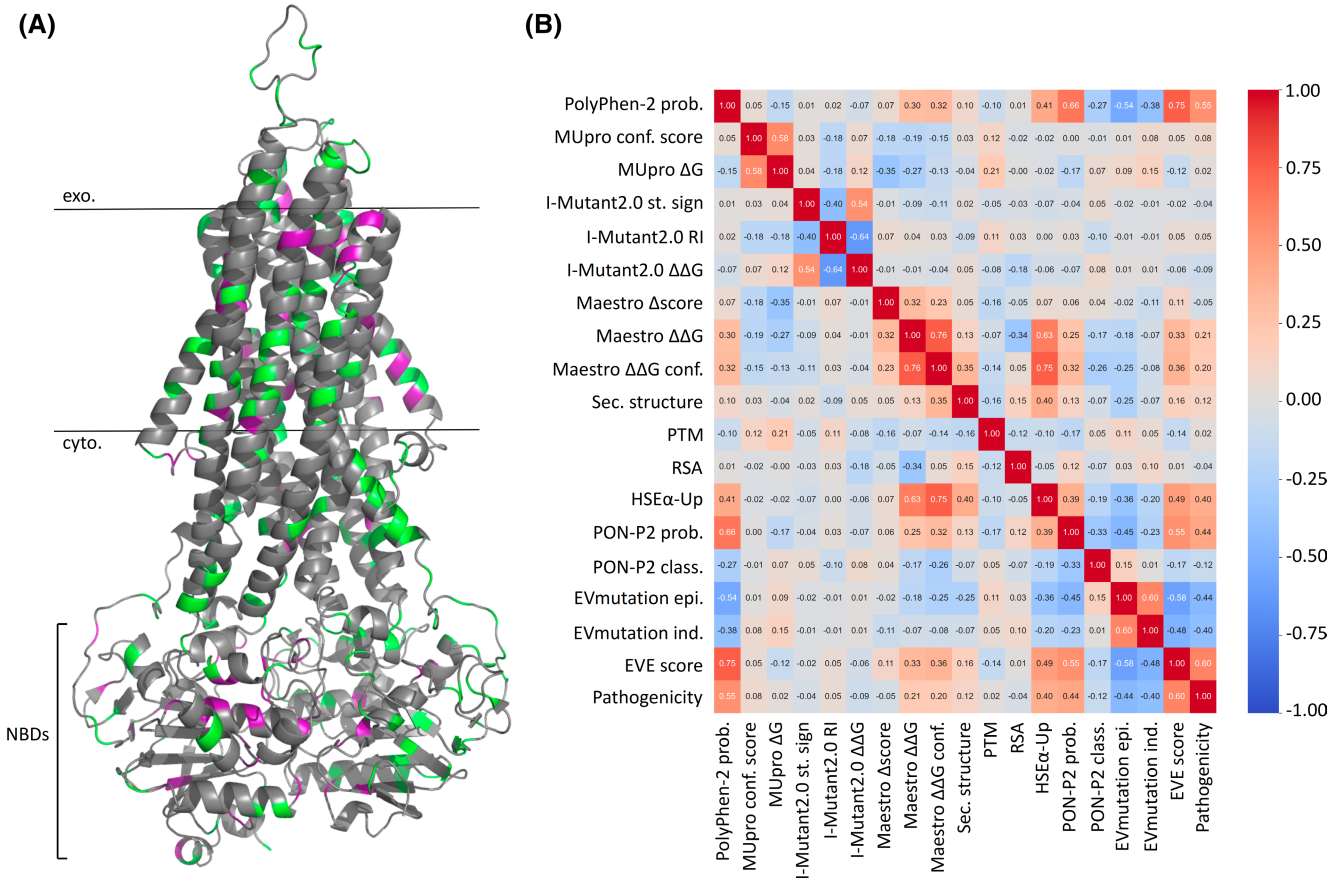

FIGURE 2 Coverage of MDR3 by the data set and correlation analysis of features. (A) Mapping of data set variants onto the MDR3 structure. Benign variants are marked in green and pathogenic variants in magenta. (B) Spearman rank correlation matrix of features computed for the data set. conf., confidence; cyto., cytosolic; epi., epistatic; EVE, Evolutionary Models of Variant Effects; exo., extracellular; HSE, half-sphere exposure; ind., independent; MDR3, multidrug resistance protein 3; NBD, nucleotide-binding domain; prob., probability; PTM, posttranslational modification; RI, reliability index; RSA, relative solvent accessibility; Sec. structure, secondary structure; st. sign, stability sign.

uncertainty of those integrated general predictors that are based on evolutionary sequence conservation. Overall, also given the small number of misclassifications, we do not see indications of domains of increased uncertainty for MDR3 predictions. The correlation coefficients between input features range from −0.64 to 0.76 (RMS value, 0.25) over the 18 features (Figure 2B), indicating that each feature adds information that does not overlap with information from another feature.

Generating Vasor: training the XGBoost algorithm on the data set

For machine-learning models to function reliably, it is vital to estimate potential overfitting or underfitting of the trained model. One of the most important techniques in that respect is the hold-out method, where a subsection of the entire data set is split off as an external test set. Ideally, the test set has a similar probability distribution as the entire data set[45]; however, this is not certain if a test set is randomly drawn. Therefore, we paid attention to drawing our test set with a similar distribution of amino acids as to both reference sequence and variant amino acid distributions by minimizing the RMSD-based difference in amino acid distributions to the overall data set; the test set contained 20 benign and pathogenic variants each (Figure S1).

Next, for the remaining data set, SMOTE[38] was used to create synthetic examples of the minority class (pathogenic variants) to balance the classes. The final training data set consisted of 259 data points for each class, benign and pathogenic, on which an XGBoost algorithm was trained. To evaluate the most important features, we measured and visualized feature importance (Figure S4) and removed the four consistently least important features (Figure S5) without reducing performance. Of note, EVE is highly important for the prediction outcome of the model, indicating that Vasor primarily relies on EVE's predictions compared to other features.

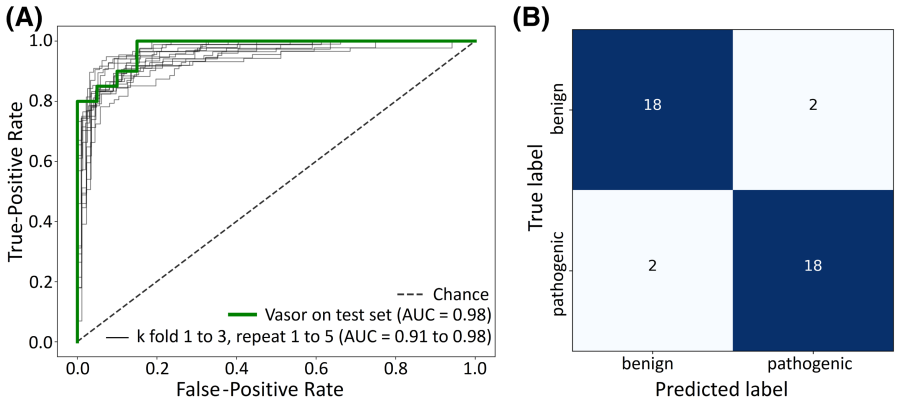

FIGURE 3 Performance of Vasor on the test set. (A) ROC curve of Vasor performance on the test set (green line) compared to performance estimates from repeated *k*-fold cross-validation (black lines). (B) Confusion matrix of Vasor performance on the test set. AUC, area under the curve; ROC, receiver operating characteristic; Vasor, variant assessor of MDR3.

Performance estimates were visualized within a repeated *k*-fold cross-validation and compared to the performance against the held-out test set (Figure 3A). The trained model performs on the test set with an accuracy of 90%, with 18 out of 20 variants being predicted correctly, both for the benign and the pathogenic class (Figure 3B). Notably, the performance based on the *k*-fold cross-validation does not differ from that on the independent test set, indicating a well-fit model without overfitting or underfitting.

Vasor outperforms integrated general predictors and the external general predictor MutPred2

We compared the performance of Vasor with general predictors on the entire data set. We compared Vasor to EVE, PolyPhen-2, and PON-P2, integrated as features into the data set on which Vasor was trained. Vasor should outperform each predictor due to the additional information gathered from the other features. Additionally, we compared Vasor to MutPred2[25] as an external prediction tool; the predecessor tool MutPred was indicated to perform well on MDR3 classification problems.[16] Vasor outperformed EVE, PolyPhen-2, PON-P2, and MutPred2 according to ROC (Figure 4A) and precision-recall curves (Figure 4C), with an area under the curve (AUC) of 0.98 for Vasor against 0.90 for EVE, 0.89 for MutPred2, 0.87 for PolyPhen2, and 0.81 for PON-P2 for the ROC and an AUC of 0.94 for Vasor against an AUC of 0.86 for EVE, 0.74 for MutPred2, 0.72 for PolyPhen2, and 0.55 for PON-P2 for the precision-recall curves. Precision-recall curves have been shown to be more robust and accurate for binary classifiers on imbalanced data sets.[46]

Noteworthy, the second best performing predictor, EVE, was the most important feature for Vasor, suggesting that the machine-learning model recognized the information contained within this feature as highly correlated with the true output and its value in predicting the output correctly. However, EVE could only predict 85.7% of the variants in the data set, whereas Vasor, by design, predicted an outcome for every possible missense variant of MDR3 (Figure 4B; Table 1).

Additional performance measures are summarized in Table 1, indicating that Vasor outperforms existing prediction tools according to the weighted measures F1-score (0.85) and MCC (0.80). Specifically, Vasor achieved a low number of FNs. Comparable low values in FNs were achieved by PolyPhen2 and MutPred2 (but at the cost of an increased number of FPs) and PON-P2, but only at coverage of 45.1% of the variants in the MDR3 protein and an increased number of FPs.

When comparing the performance of the missense predictors on the test set (Table S2), our tool reached the best scores in F1-score and MCC (0.90 and 0.80, respectively) compared to other predictors with full coverage of the test set. EVE showed F1-score and MCC values of 0.91 and 0.83, respectively, on a subset (82.5%) of variants where it reached a prediction. By contrast, MutPred2 was able to predict every pathogenic variant as pathogenic, albeit at the cost of predicting almost half of the benign variants as pathogenic, resulting in a high number of FPs.

Overall, Vasor outperformed other predictors consistently according to ROC and precision-recall curves, revealing a well-balanced prediction with few FNs and FPs, both on the entire data set and the test set.

Vasor classifies the majority of variants with high certainty

Additionally, we investigated the distribution of Vasor's output, the PoP values. Vasor assigns the majority of benign cases low probability values (74% of benign variants <0.24 PoP), whereas the majority of

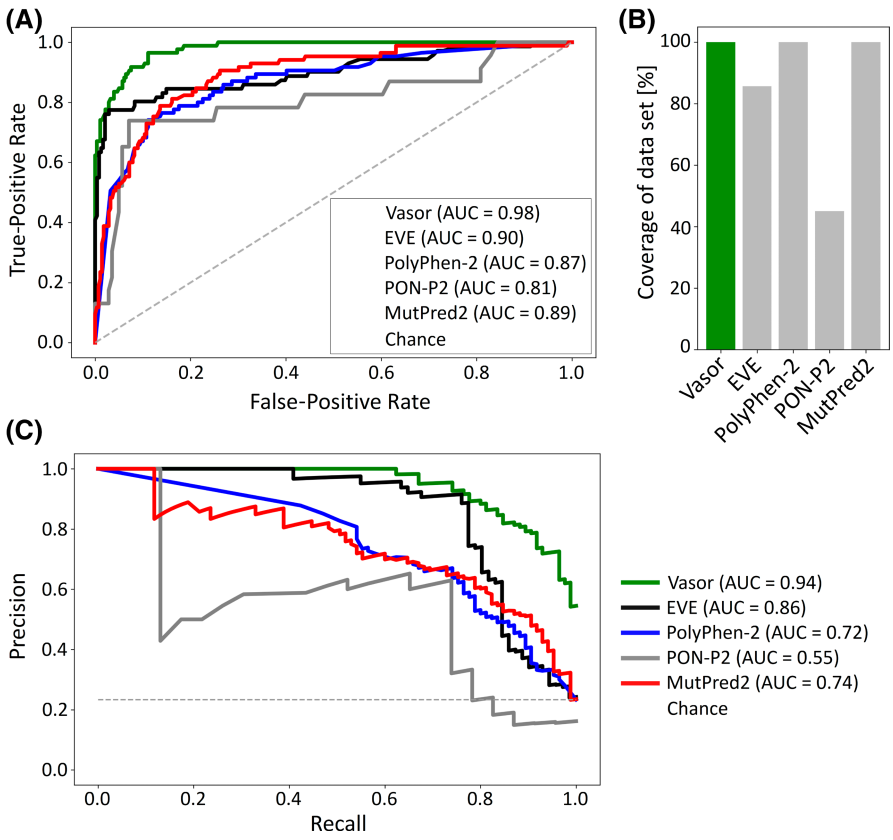

FIGURE 4 Performance of Vasor in comparison to established general predictors. (A) ROC curve of the performance of Vasor, EVE, PolyPhen-2, PON-P2, and MutPred2 on the variants of the entire data set. Note that the performance was determined for those variants each predictor was able to make a prediction for (see [B]). (B) Coverage of data set variants by the predictors. (C) Precision-recall curves of the predictors. Performance was determined for those variants each predictor was able to make a prediction for. AUC, area under the curve; EVE, Evolutionary Models of Variant Effects; ROC, receiver operating characteristic; Vasor, variant assessor of MDR3.

TABLE 1 Detailed performance measurements of Vasor in comparison to EVE, PolyPhen-2, PON-P2, and MutPred2 on the entire data set

	Vasor	EVE	PolyPhen-2	PON-P2	MutPred2
Recall	0.84	0.73	0.84	0.74	0.93
Specificity	0.96	0.98	0.74	0.89	0.67
Precision	0.86	0.91	0.49	0.52	0.46
NPV	0.95	0.93	0.94	0.95	0.97
Accuracy	0.93	0.92	0.76	0.87	0.73
F1-score	0.85	0.81	0.62	0.61	0.61
MCC	0.80	0.77	0.50	0.54	0.51
TP	71	52	71	17	79
FN	14	19	14	6	6
TN	267	236	206	125	186
FP	12	5	73	16	93
Coverage (%)	100	85.7	100	45.1	100

Abbreviations: EVE, Evolutionary Models of Variant Effects; FN, false negative; FP, false positive; MCC, Matthew's correlation coefficient; NPV, negative predictive value; TN, true negative; TP, true positive; Vasor, variant assessor of MDR3.

pathogenic cases are assigned a high probability value (75% of pathogenic variants >0.74 PoP) (Figure 5). Furthermore, Vasor showed no misclassifications of variants in the data set for values below 0.23 and above 0.84, indicating high certainty for benign variant predictions in the range 0–0.23 (74% of the benign variants) and pathogenic variant predictions in the range 0.84–1 (60% of the pathogenic variants).

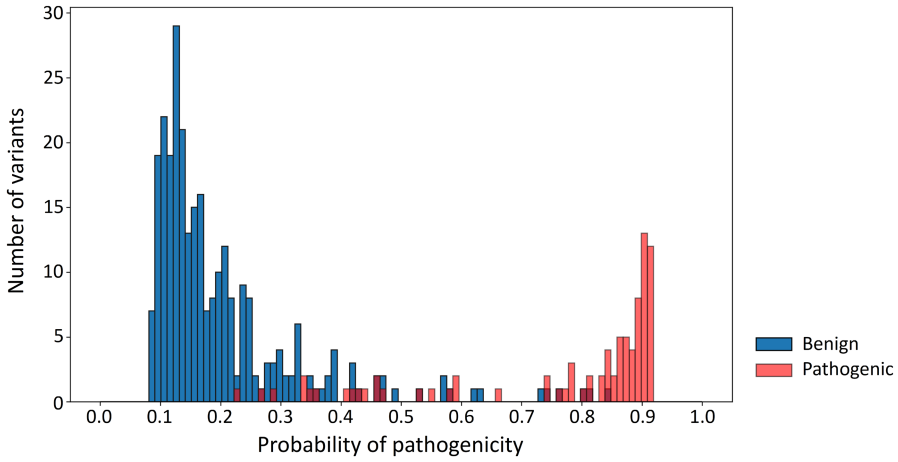

FIGURE 5 Distribution of probability of pathogenicity values over the entire data set. Distribution of Vasor's probability of pathogenicity output for benign (blue) and pathogenic (red) variants. Vasor classified 74% of benign variants into the benign category with values below 0.22, which is below the lowest probability value of any pathogenic variant (0.23) within the data set; 60% of pathogenic variants were classified into the pathogenic category with values above 0.85, which is greater than the highest probability value of any benign variant (0.84) within the data set; 75% of pathogenic variants were classified with probability values greater than 0.74. Vasor, variant assessor of MDR3.

We further investigated the use of SMOTE to generate data points for the minority class (i.e., pathogenic variants). Due to the method underlying SMOTE, SMOTE-generated data points are expected to follow the distribution of pathogenic variants within the PoP curve. Accordingly, no SMOTE data point was predicted with a lower value of PoP than 0.28, and data points mainly clustered within the high certainty zone (Figure S6).

Overall, Vasor showed a robust separation of PoP values of both variant classes, indicating that Vasor classified most variants within the data set with high certainty.

Easy accessibility of Vasor as a webserver tool

Using Vasor, we precalculated the effect of every possible amino acid substitution for MDR3, resulting in a heatmap of 1286 × 20 probabilities of pathogenicity (Figure 6; Table S3). We mapped the average PoP of each position onto the MDR3 protein structure to visualize positions that are functionally more sensitive to substitutions (Figure 7). As expected, areas near the ATP-binding site within the nucleotide-binding domain displayed a high average PoP. Similarly, buried residues within the helices forming the TM part showed high sensitivity as several missense mutations may lead to a disruption of the helical structure. More exposed residues located on the outsides of helices or in flexible regions, such as the small extracellular loops, displayed less sensitivity. However, this trend does not exclude that specific variants at seemingly less sensitive sites can be pathogenic and vice versa.

To indicate the usage of the webserver more specifically, we exemplarily predicted the effect of two variants, V428D and N902D, identified in Dröge et al.[9] These variants were identified in patients without further *in vitro* analysis and not used in the data set for creating Vasor. The variant V428D is predicted to be pathogenic by Vasor with a PoP of 0.77, indicating a good level of certainty for a correct prediction of the pathogenic effect as only four out of 12 variants from the data set were falsely predicted with a similarly high score (Figure 5). V428D is located directly before the Walker A motif, which is important for correctly coordinating the adenosine and the phosphate moiety of ATP in combination with the Walker B motif. Accordingly, the variant might disturb this recognition, resulting in a distorted functionality of MDR3. The variant N902D is predicted to be pathogenic by Vasor with a PoP of 0.90, indicating a high level of certainty for a correct prediction as no false predictions within the data set were observed at such high values (Figure 5). N902D is located in the cytosol-facing part of TM10, with the potential to interact with residues of the X loop of nucleotide-binding domain 1, especially R529. As the X loop is likely involved in relaying the ATP-binding event to the TM domains through conformational change,[47] N902D might exert its effect by hindering this transmission.

We also used the precomputed heatmap for rapid lookup and output generation of the webserver tool, thus eliminating waiting time for users needing a prediction for a specific MDR3 variant. The webserver can be accessed at https://cpclab.uni-duesseldorf.de/mdr3_predictor/. It requires as input an MDR3 variant (with the amino acid of the reference sequence in the one-letter format, its position within the canonical sequence of Uniprot ID: P21439-1, and the substituted amino acid

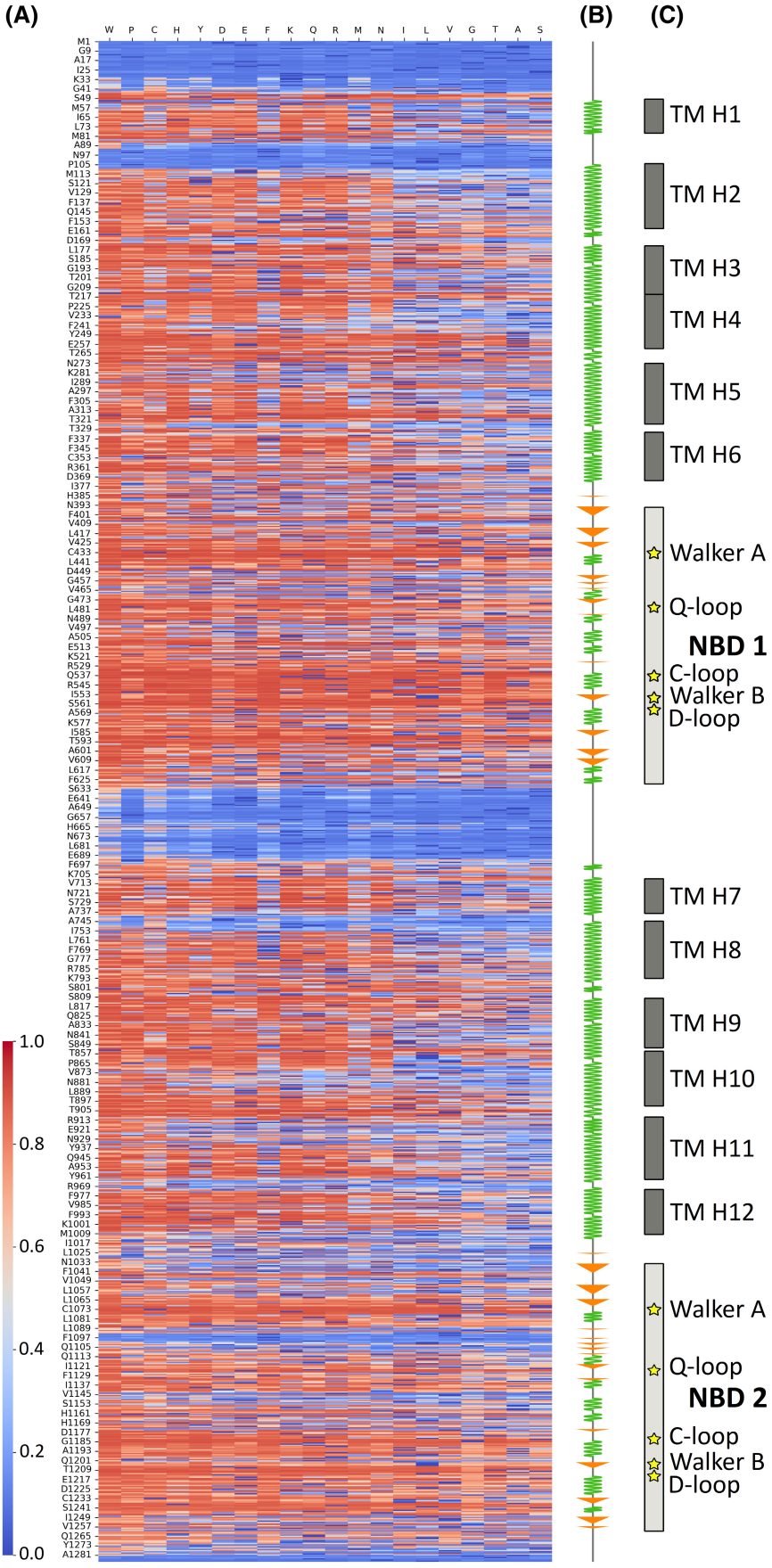

FIGURE 6 Heatmap of predictions for every possible amino acid substitution in MDR3. (A) Color-coded predictions for every position (displayed on the *y* axis) within the MDR3 protein and every possible amino acid substitution (*x* axis). Prediction values range from likely benign (blue) to likely pathogenic (red). (B) Secondary structure of MDR3. α-helical stretches are depicted as green zig-zag curves, β-sheet stretches as orange arrows. (C) Domains, secondary structure elements, and characteristic motives are indicated on the right. MDR3, multidrug resistance protein 3; NBD, nucleotide-binding domain; TM H, transmembrane helix.

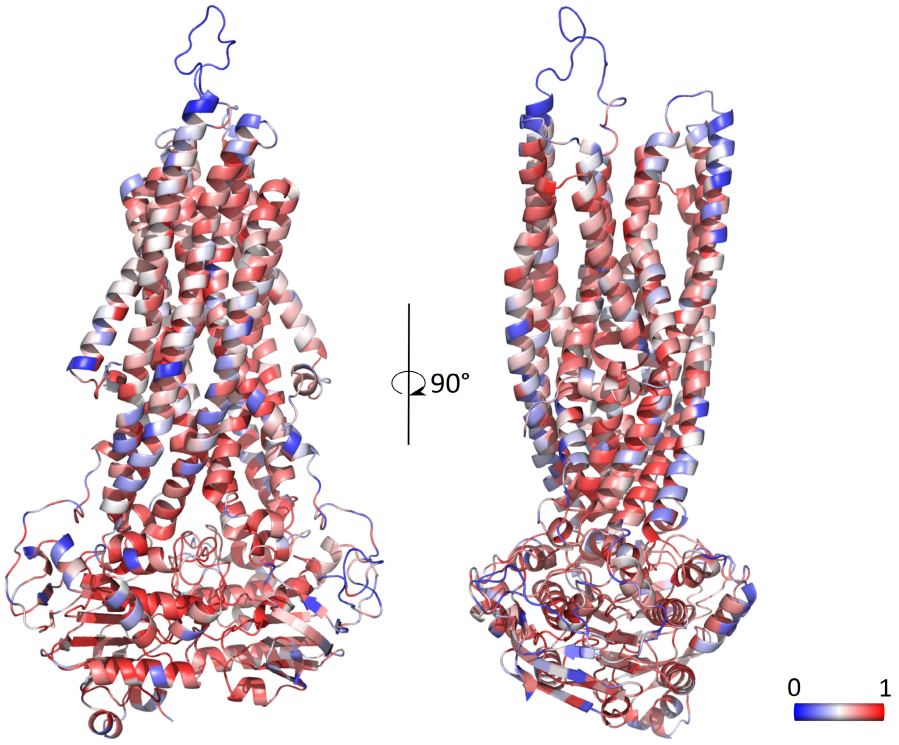

FIGURE 7 Mapping the average pathogenicity onto the structure of MDR3. Prediction values for each position were averaged over all possible substitutions. Values closer to 0 (most likely benign) correspond to blue, values closer to 1 (most likely pathogenic) correspond to red residues. MDR3, multidrug resistance protein 3.

in the one-letter format) and yields the predicted effect of the entered variant, either benign or pathogenic, together with the PoP. Additionally, the variant position is depicted in the three-dimensional structure of MDR3, and high-quality images of reference sequence amino acid, variant, and the overall MDR3 structure can be downloaded. The heatmap is also downloadable from the webserver for implementation in other applications.

DISCUSSION

Although recent years have resulted in many general predictors for protein properties, their performance on specific proteins of interest can differ greatly.[48] While existing state-of-the-art tools to predict substitution effects perform admirably on the MDR3 protein, especially EVE,[18] the potential for improvement is given both for the performance on and coverage of the MDR3 data set because not every general predictor can classify each MDR3 variant. To improve predictions, we created what is to our knowledge the largest data set specific for pathogenic and benign variants of MDR3, obtained from the literature and gnomAD database and comprising 85 pathogenic and 279 benign variants. As the generation of a high-quality data set is a critical first step for any machine-learning approach,[45,49] we carefully screened the literature specifically for MDR3 variants, filtering out variants with unclear disease associations. To counteract the bias that mainly pathogenic variants are chosen for detailed *in vitro* or *in vivo* analysis, we obtained variants from the gnomAD database.[26] Because there may be potentially disease-associated variants in the database, we implemented an additional filtering step of removing variants categorized as likely pathogenic or pathogenic as evaluated by VarSome[27] to exclude FN variants. The data set resulting from this strategy was then kept as is (i.e., no variants were added or removed), thus eliminating the potential to introduce bias from the researcher. Using established general predictors and variant site properties, we trained an MDR3-specific machine-learning model, termed Vasor, to classify

protein missense variants into benign or pathogenic. Vasor outperformed general predictors. Over the entire data set, Vasor showed F1-score and MCC values of 0.85 and 0.80, respectively; the second best method, EVE, followed with scores of 0.81 and 0.77, respectively, but coverage of only 85.7%. By contrast, Vasor ensured high-quality predictions for all MDR3 missense variants. As machine-learning models trained on a specific data set exhibit a bias toward overperformance on this data set, Vasor has an inherent advantage when evaluated on the entire data set over other predictors. Notably, the superior performance of Vasor was also present on the independent test set where Vasor only misclassified two (5%) benign and two (5%) pathogenic variants, leading to the highest performance compared to other predictors, as indicated by F1-score and MCC of 0.9 and 0.8, respectively. Although EVE and PON-P2 achieved similar performances for the test set, they only covered a fraction of the variants (82.5% and 37.5%, respectively). Overall, no other analyzed predictor provided a similarly good balance of consistently low FN and FP predictions. Both measures have important implications for using Vasor within a clinical setting. Predictors with a high number of FNs will lead to variants found within patients being falsely given no attention, whereas a high number of FPs will result in a predictor raising too often a false alarm for an actually benign variant.

We established an easily accessible webserver for reliable and fast predictions of novel MDR3 variants based on Vasor. It can serve as an important step for deciding which variants to study and to provide the first indication of a variant effect. It does not eliminate the need for classical *in vitro* studies for mutational impact, however, and in a clinical setting, the ACMG-AMP guidelines[28] should be followed. The webserver classifies single-site amino acid substitutions into the categories benign or pathogenic. Truncation, insertion, and deletion variants of MDR3 cannot be assessed. However, the PoP for such variants is often more definite.[50] Of note, the effect of a single missense variant within the biological context might not always be a clear-cut pathogenic or benign effect. Therefore, the PoP provided by the webserver can act as an indicator of prediction reliability.

As a limitation, the exact mechanism underlying a pathogenic variant cannot be inferred from the current tool. MDR3 missense variants may impact protein folding and maturation, activity, or stability,[13] and several of these categories can be influenced. Information on mechanistic dysfunction may aid in targeted therapy. In terms of machine learning, such a multiclass classification problem might be solved—with the premise of a sizeable data set of quality-assured variants. Unfortunately, we are unaware of such a data set for MDR3. The currently employed data set strived for such quality-assured variants; however, especially lacking large-scale functional studies of benign variants, variants indicated by VarSome as of unclear significance were included. Thus, we encourage the scientific community to submit novel MDR3 variants with a proven effect on folding, maturation, activity, and stability to the authors to be added to the data set to improve and develop Vasor further.

ACKNOWLEDGMENTS

We are grateful for computational support and infrastructure provided by the "Zentrum für Informations-und Medientechnologie" at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing to Holger Gohlke on the supercomputer JUWELS at Jülich Supercomputing Centre (user ID: HKF7, VSK33, FIC). Open Access funding enabled and organized by Projekt DEAL. WOA Institution: HEINRICH-HEINE-UNIVERSITAET DUESSELDORF Consortia Name: Projekt DEAL.

FUNDING INFORMATION

BMBF through HiChol (translational network on hereditary intrahepatic cholestasis); Grant Numbers: 01GM1904A, 01GM1904B

CONFLICT OF INTEREST

Verena Keitel is on the speakers' bureau of Falk Foundation and Albireo and advises Astra Zeneca. The other authors have nothing to report.

ORCID

Holger Gohlke https://orcid.org/0000-0001-8613-1447

REFERENCES

1. Smith AJ, Timmermans-Hereijgers JL, Roelofsen B, Wirtz KW, van Blitterswijk WJ, Smit JJ, et al. The human MDR3 P-glycoprotein promotes translocation of phosphatidylcholine through the plasma membrane of fibroblasts from transgenic mice. FEBS Lett. 1994;354(3):263–6.
2. van Helvoort A, Smith AJ, Sprong H, Fritzsche I, Schinkel AH, Borst P, et al. MDR1 P-glycoprotein is a lipid translocase of broad specificity, while MDR3 P-glycoprotein specifically translocates phosphatidylcholine. Cell. 1996;87(3):507–17.
3. Oude Elferink RPJ, Paulusma CC. Function and pathophysiological importance of ABCB4 (MDR3 P-glycoprotein). Pflugers Arch. 2007;453:601–10.
4. Olsen JA, Alam A, Kowal J, Stieger B, Locher KP. Structure of the human lipid exporter ABCB4 in a lipid environment. Nat Struct Mol Biol. 2020;27(1):62–70.
5. Prescher M, Bonus M, Stindt J, Keitel-Anselmino V, Smits SHJ, Gohlke H, et al. Evidence for a credit-card-swipe mechanism in the human PC floppase ABCB4. Structure. 2021;29(10):1144–55.e5.
6. Rosmorduc O, Hermelin B, Poupon R. MDR3 gene defect in adults with symptomatic intrahepatic and gallbladder cholesterol cholelithiasis. Gastroenterology. 2001;120(6):1459–67.
7. Deleuze J, Jacquemin E, Dubuisson C, Cresteil D, Dumont M, Erlinger S, et al. Defect of multidrug-resistance 3 gene expression in a subtype of progressive familial intrahepatic cholestasis. Hepatology. 1996;23(4):904–8.
8. Lang C, Meier Y, Stieger B, Beuers U, Lang T, Kerb R, et al. Mutations and polymorphisms in the bile salt export pump

and the multidrug resistance protein 3 associated with drug-induced liver injury. Pharmacogenet Genomics. 2007; 17(1):47–60.
9. Dröge C, Bonus M, Baumann U, Klindt C, Lainka E, Kathemann S, et al. Sequencing of FIC1, BSEP and MDR3 in a large cohort of patients with cholestasis revealed a high number of different genetic variants. J Hepatol. 2017;67(6):1253–64.
10. Pauli-Magnus C, Lang T, Meier Y, Zodan-Marin T, Jung D, Breymann C, et al. Sequence analysis of bile salt export pump (ABCB11) and multidrug resistance p-glycoprotein 3 (ABCB4, MDR3) in patients with intrahepatic cholestasis of pregnancy. Pharmacogenetics. 2004;14:91–102.
11. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435–44.
12. Dong C, Condat B, Picon-Coste M, Chrétien Y, Potier P, Noblinski B, et al. Low-phospholipid-associated cholelithiasis syndrome: prevalence, clinical features, and comorbidities. JHEP Rep. 2020;3(2):100201.
13. Delaunay JL, Durand-Schneider AM, Dossier C, Falguières T, Gautherot J, Davit-Spraul A, et al. A functional classification of ABCB4 variations causing progressive familial intrahepatic cholestasis type 3. Hepatology. 2016;63(5):1620–31.
14. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M. A review study: computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene. 2019;680:20–33.
15. Niroula A, Vihinen M. Variation interpretation predictors: principles, types, performance, and choice. Hum Mutat. 2016;37:579–97.
16. Khabou B, Durand-Schneider AM, Delaunay JL, Aït-Slimane T, Barbu V, Fakhfakh F, et al. Comparison of in silico prediction and experimental assessment of ABCB4 variants identified in patients with biliary diseases. Int J Biochem Cell Biol. 2017;89:101–9.
17. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.
18. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–5. Erratum in: Nature. 2022;601(7892):E7.
19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
20. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33:W306–10.
21. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006;62(4):1125–32.
22. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics. 2015;16:116.
23. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction method for fast and reliable identification of harmful variants. PLoS One. 2015;10(2):e0117380.
24. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128–35.
25. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11(1):5918.
26. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al.; Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. Erratum in: Nature. 2021;590(7846):E53.
27. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.
28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24.
29. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
30. Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. Nat Biotechnol. 2007;25(3):285–6.
31. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43:D512–20.
32. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294(5):1351–62.
33. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48:D296–306.
34. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39:D411–9.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
36. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PLoS One. 2013;8(11):e80635.
37. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins. 2005;59(1):38–48.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
39. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Available from: https://arxiv.org/abs/1603.02754. Accessed March 1, 2021.
40. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13(Suppl 4):S2.
41. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018;34:3755–8.
42. Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 2015;43(W1):W576–9.
43. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012;40:D370–6.
44. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17(2):109–21.
45. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. 2018. Available from: https://arxiv.org/abs/1811.12808. Accessed May 15, 2021.
46. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
47. Dawson RJP, Locher KP. Structure of a bacterial multidrug ABC transporter. Nature. 2006;443(7108):180–5.
48. Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat. 2016;37(10):1013–24.
49. Walsh I, Pollastri G, Tosatto SCE. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016;17(5):831–40.

50. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Behrendt A, Golchin P, König F, Mulnaes D, Stalke A, Dröge C, et al. Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3. Hepatol Commun. 2022;6:3098–3111. https://doi.org/10.1002/hep4.2088