Improved *KCNQ2* gene missense variant interpretation with artificial intelligence

Alba Saez-Matia[1*], Arantza Muguruza-Montero[1], Sara M-Alicante[1,2], Eider Núñez[1,2], Rafael Ramis[2,4], Óscar R. Ballesteros[2,3], Markel G Ibarluzea[2], Carmen Fons[5], Aritz Leonardo[2,4], Aitor Bergara[2,3,4] and Alvaro Villarroel[1*]

[1]Instituto Biofisika, CSIC-UPV/EHU, 48940 Leioa, Spain
[2]Departamento de Física, Universidad del País Vasco, UPV/EHU, 48940 Leioa, Spain
[3]Centro de Física de Materiales CFM, CSIC-UPV/EHU, 20018 Donostia, Spain
[4]Donostia International Physics Center, 20018 Donostia, Spain
[5]Pediatric Neurology Department, Sant Joan de Déu Hospital, Barcelona University, Barcelona, Spain

[*]To whom correspondence should be addressed at alvaro.villarroel@csic.es or alba.saezmatia@gmail.com

ABSTRACT

Advances in DNA sequencing technologies have revolutionized rare disease diagnosis, resulting in an increasing volume of available genomic data. Despite this wealth of information and improved procedures to combine data from various sources, identifying the pathogenic causal variants and distinguishing between severe and benign variants remains a key challenge. Mutations in the K_v7.2 voltage-gated potassium channel gene (*KCNQ2*) have been linked to different subtypes of epilepsies, such as benign familial neonatal epilepsy (BFNE) and epileptic encephalopathy (EE). To date, there is a wide variety of genome-wide computational tools aiming at predicting the pathogenicity of variants. However, previous reports suggest that these genome-wide tools have limited applicability to the *KCNQ2* gene related diseases due to overestimation of deleterious mutations and failure to correctly identify benign variants, being, therefore, of limited use in clinical practice. In this work, we found that combining readily available features, such as AlphaFold structural information, Missense Tolerance Ratio (MTR) and other commonly used protein descriptors, provides foundations to build reliable gene-specific machine learning ensemble models. Here, we present a transferable methodology able to accurately predict the pathogenicity of *KCNQ2* missense variants with unprecedented sensitivity and specificity scores above 90%.

INTRODUCTION

Over the past decade, the rapid progress in genomic technologies has made research and clinical sequencing increasingly accessible [Traynelis *et al.,* 2017]. However, identifying the true phenotypically causal variants remains a major challenge, mainly in terms of time and costs [McInnes *et al.,* 2021; Finn *et al.*, 2009]. In addition, weak effect sizes and correlations among neighboring variants, locus heterogeneity and limitations in sample or pedigree sizes limit the discovery power and resolution of purely genetic studies [Cooper & Shendure, 2011]. This problem is exacerbated for rare diseases and complex pathologies where genetic information by itself is often insufficient [McInnes *et al.,* 2021; Mayer *et al.,* 2011; MacArthur & Smith, 2010]. This scenario is especially visible for *KCNQ2* gene related pathologies.

KCNQ2 gene encodes the voltage-gated potassium channel subunit $K_v7.2$, which plays a crucial role in neuronal excitability by mediating the M-current [Etxeberrria *et al.,* 2008; Schwake *et al.,* 2006; Gutman *et al.*, 2003; Marrion, 1997]. Initially, autosomal dominantly inherited *KCNQ2* mutations were linked to benign familial neonatal epilepsy (BFNE) [Singh *et al.,* 1998; Biervert *et al.,* 1998]. Later on, *de novo KCNQ2* mutations were identified in patients with neonatal developmental and epileptic encephalopathy (DEE, EE) [Goto *et al.,* 2021; Millichap *et al.,* 2016; Pisano *et al.,* 2015; Numis *et al.,* 2014; Kato *et al.,* 2013; Weckhuysen *et al.*, 2013; Weckhuysen *et al.,* 2012]. *KCNQ2* encephalopathy presents early neonatal onset intractable seizures, usually tonic seizures with suppression burst EEG pattern and severe developmental delay [Millichap *et al.,* 2016; Pisano *et al.,* 2015; Kato *et al.,* 2013; Weckhuysen *et al.*, 2013]. Intractable seizures and poor prognosis warrant an urgent need for effective early diagnosis and treatment [Kim *et al.,* 2021]. We strongly believe that this effective treatment must start with a correct identification of *KCNQ2* pathogenic variants.

To achieve that aim, the use of computational tools has been extended to many fields of medical sciences, including development of predictive software to assess the pathogenicity of protein variants [Wainberg *et al.,* 2018; Thusberg *et al.,* 2011]. To date, there are a wide variety of computational tools aiming at predicting the pathogenicity of variants [Katsonis *et al.,* 2021]. Some commonly used tools include Sorting Intolerant From Tolerant (SIFT, [Ng, 2003]), Protein Variation Effect Analyzer (PROVEAN, [Choi *et al.,* 2012]) and Functional Analysis through Hidden Markov Models (FATHMM, [Shihab *et al.,* 2013]), which use a sequence homology-based method; Polymorphism Phenotyping v2 (PolyPhen2, [Adzhubei *et al.,* 2010]), which utilizes a sequence and structure-based approach; MutPred2 [Pejaver *et al.,* 2020], which combines evolutionary, structural and function-based features; and consensus deleteriousness score (CONDEL, [González-Pérez & López-Bigas, 2011]) which uses a consensus-based approach. Despite relative success in differentiating between severe disease-causing and benign variants in some genes, these tools are not consistently effective in their predictions and their accuracy has been shown to be gene-dependent [Leong *et al.,* 2015; Chun & Fay, 2009]. Consequently, there is no *in silico* tool capable of effectively predicting the pathogenicity of *KCNQ2* variants to date with sensitivity/specificity scores above 90%, which characterize a clinical relevant tool [Richards *et al.,* 2015]. Nevertheless, machine learning-based techniques have already proven their utility for other diseases related to genes encoding ion channels [Xenakis *et al.,* 2021, Sallah *et al.,* 2020, Li *et al.,* 2017].

Machine learning (ML) refers broadly to the process of fitting predictive models to data or of identifying informative groupings within data. The field of ML essentially attempts to approximate or imitate the ability to recognize patterns of humans, albeit in an objective manner, through algorithms and computation [Greener *et al.,* 2021]. Supervised ML refers to techniques in which a model is trained on a range of inputs (or features) which are associated with a known outcome. In medicine, this might represent training a model to relate the characteristics of an individual (e.g., height, weight, smoking status) to a certain outcome (onset of diabetes within five years, for example). Once the algorithm is successfully trained, it should be capable of making outcome predictions when applied to new data [Jenni *et al.,* 2019].

It is accepted that the function of proteins emanates from the structure, which in turn depends on its amino acid sequence. However, the use of 3D information has been hampered by the lack of resolved

structures at atomic resolution, especially for membrane proteins such as ion channels. The cryo-EM revolution has ameliorated this limitation to great extent, but most of the published structures only provide partial coverage (i.e., some large stretches of amino acids are missing in the coordinate files). For instance, the recent $K_V7.2$ structure submitted in the RCSB Protein Data Bank (https://www.rcsb.org/) (PDB: 7CR3) covers only 40% of the total amino acid sequence. Fortunately, the release of AlphaFold2 [Jumper *et al.,* 2021] opened up a new era in the understanding of protein structure. The deep neural network of this algorithm, which combines features derived from homologous templates and from multiple sequence alignment to generate the predicted structure, has shown an outstanding accuracy in predicting the three-dimensional structure of proteins with otherwise unknown fold [David *et al.,* 2022]. In this study, we investigated methods to capture this structural information to help discriminating the pathogenic consequences of missense mutations. Combining these with more traditional descriptors, we developed the first *KCNQ2*-specific machine learning-ensemble algorithm (MLe-KCNQ2) capable of accurately predicting the pathogenicity of channel $K_V7.2$ variants. This novel algorithm, with a sensitivity of 91.75% and a specificity of 90.21%, shows a high accuracy at predicting both pathogenic and benign variants, respectively. Therefore, MLe-KCNQ2 could help clinicians and researchers to interpret the effect of missense mutations in the *KCNQ2* gene as well as facilitate the clinical diagnosis and early choice of personalized therapies for developmental *KCNQ2*-related disorders. The methodology employed here can be extended to the prediction of the pathogenicity of variants occurring in other genes.

METHODS

DATASET

Isoform 4 of the *KCNQ2* gene encodes a transcript with 872 codons, which could lead to 7.582 single base substitutions, leading to 5.972 single amino acid missense variants (excluding stop codons). To date, more than 1.200 *KCNQ2* variants have been annotated in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) (16/09/21). These variants are divided into 6 subclasses according to the type of the mutation: frameshift, missense, nonsense, splice site, non-coding RNA, and untranslated region (UTR). In order to develop an efficient classification algorithm, three aspects must be taken into account: (i) each subclass must be analyzed individually because a given feature could be important for one subclass but not relevant for others [Larrea *et al.,* 2021]; the chosen subclass must have (ii) an adequate number of variants and (iii) at least two clinical conditions with distinct effect (e.g., benign and pathogenic significance). This study focuses on missense variants because they are the only ones that meet all these criteria.

More than 500 missense variants are already annotated in ClinVar, which are distributed according to their pathogenicity as follows: 6 benign, 12 likely benign, 254 of uncertain significance, 39 of conflictive interpretation, 122 likely pathogenic and 141 pathogenic. It should be noted that some of these variants have been included into more than one subclass. In order to simplify the classification as a binary problem, benign and likely benign variants were grouped into the benign class, and pathogenic and likely pathogenic variants into the pathogenic class. At this point, uncertain significance and conflicting interpretation variants were isolated. As a result, an initial dataset consisting of 12 benign and 240 pathogenic variants was obtained.

In addition, an exhaustive database and bibliographic search was carried out. This procedure allowed us to incorporate missense variants not present in the initial dataset. If a variant was shared between two databases, the most recent entry was chosen. As a result, the number of variants with a reliable diagnosis was increased. In sum, the final dataset consisted of 39 benign and 314 pathogenic variants distributed along all $K_v7.2$ functional domains (Figure 1). After that, clinical labels were binarised, with benign and pathogenic variants equal to 0 and 1, respectively.

VARIANTS CHARACTERIZATION

The final dataset was characterized using 16 features. These characteristics can be divided into 3 different subgroups: features related to physico-chemical changes (amino acid charge, hydrophobicity, molecular weight, volume, polarity, aromaticity and amino acid mean solvent accessibility), biological-evolutive features (evolutive conservation value across species of the substituted residue, original and mutated residue, topological and functional affected domains, and Missense Tolerance Ratio or MTR score [Traynelis *et al.,* 2017]); and structural features (secondary structure affected by the mutation, AlphaFold2 pLDDT score and location within the structural landscape of *KCNQ2*).

MACHINE LEARNING ALGORITHMS

The different procedures shown below were performed using Python (v3.8.8) [Van Rossum & Drake, 2009] and the pandas (v1.2.4) [McKinney, 2010], NumPy (v1.20.1) [Harris *et al.,* 2020] and scikit-learn (v1.1.1) [Pedregosa *et al.,* 2011] libraries. Visualizations were performed using the Matplotlib library (v3.3.4) [Hunter, 2007].

DATA PREPROCESSING

The final dataset was subjected to format homogenization, control of outliers and missing values as well as elimination of duplicates. As a result, 9 pathogenic variants were finally discarded. The remaining 344 variants were randomly assigned to training (75%) and test (25%) sets. The training and test sets consisted of 258 (31 benign and 227 pathogenic) and 86 variants (8 benign and 78 pathogenic), respectively. As the dataset was clearly unbalanced, an oversampling technique was applied in the training set in order to artificially increase the number of benign variants. Oversampling was performed by increasing the amount of minority class samples by producing new instances or repeating some of them [Mohammed *et al.,* 2020]. In sum, 113 benign variants were included. After that, categorical features (e.g. amino acid charge change or topological affected domains) were encoded. First, these categorical values were transformed into numerical labels using the scikit-learn LabelEncoder class. After that, a one-hot encoding scheme was applied (scikit-learn OneHotEncoder class).

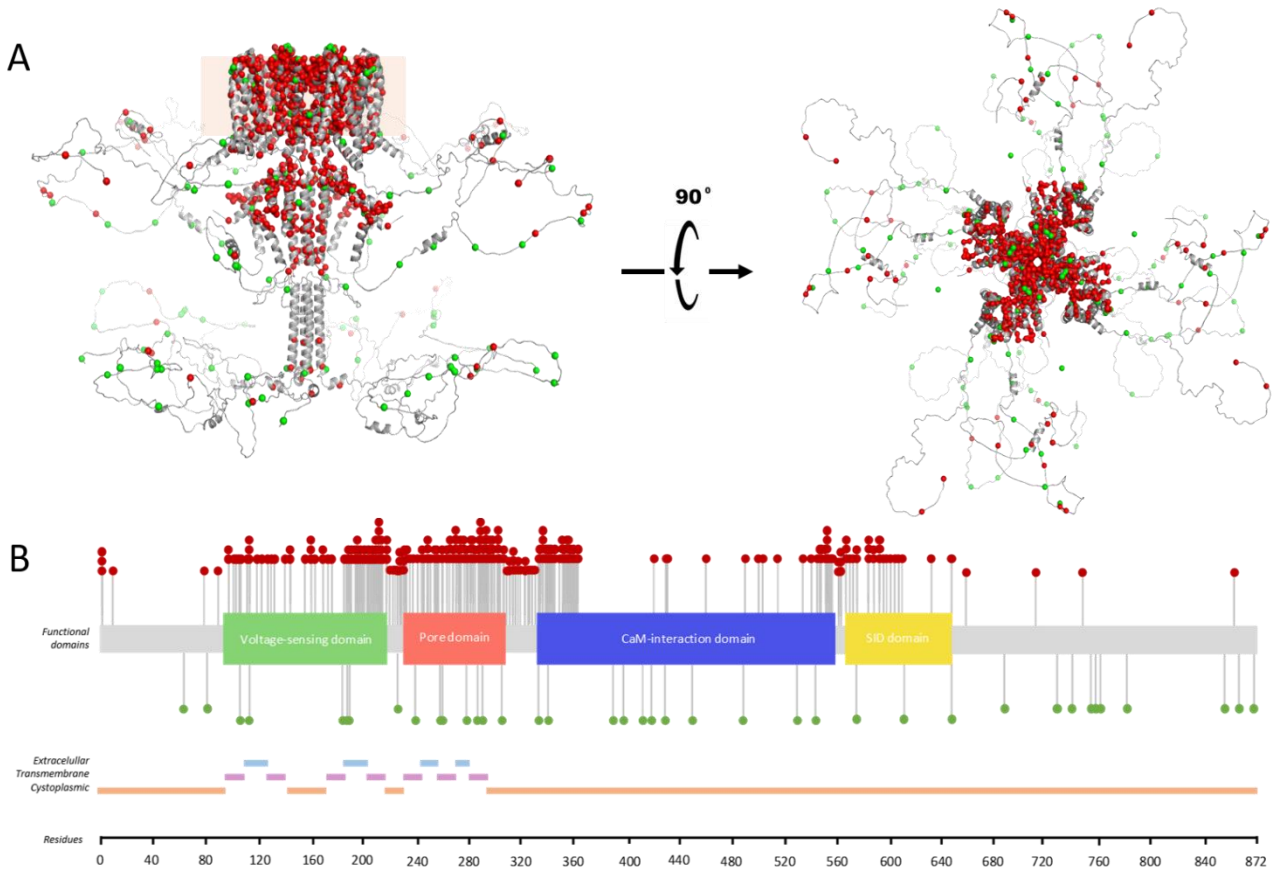

Figure 1. Graphical distribution of the 353 collected variants. (A) Distribution of the collected variants on the three-dimensional structure of a $K_v7.2$ channel tetramer, lateral (left) and upper view (right). The structure of the $K_v7.2$ monomer was built using the best prediction obtained with AlphaFold2. Intrinsically Disordered Regions (IDRs) were modeled using SwissModel [Waterhouse *et al.,* 2018] to avoid wrong location of these regions in the membrane and clashes after tetramerization. Both figures were created using PyMol [DeLano, 2002] and VMD [Humphrey *et al.*, 1996]. Pathogenic and benign variants are colored in red and green, respectively, and the membrane is indicated with an orange square in the lateral view. **(B)** Distribution of the collected variants along the main functional domains of the channel. Both the 305 pathogenic variants and the 39 benign variants are distributed throughout the entire $K_V7.2$ sequence using the previous color code. In addition to being predominant, pathogenic variants are found most abundantly in the voltage-sensing domain (VSD), the pore and the first residues of the CaM-interaction domain. In contrast, the collected benign variants are homogeneously distributed. Information about extracellular (light blue), transmembrane (pink) and cytoplasmic (orange) segments is also shown.

MODEL DEFINITION, TRAINING AND OPTIMIZATION

In a first approach, three simple ML algorithms were trained: Ridge Regression (i.e., a linear regression model that implements L2 norm for regularization) (LR2), Support Vector Classifier (SVC) and

Random Forest (RF). Previous studies have shown that these models are useful in variant pathogenicity prediction and also in unbalanced class scenarios [Greener *et al.,* 2021]. Each algorithm was trained individually using a pipeline with data standardization and the classifier. A feature selection method based on F-score (ANOVA) was applied during training. This feature selection method is used to select a subset of relevant features for effective classification, i.e., to automatically identify the most informative features out of the original ones and use them in the classification problem [Hoque *et al.,* 2014, Polat *et al.,* 2009]. Hyperparameters of each model were adjusted using a Grid Search algorithm on a 10 times 10-fold cross validation. A random seed was used to ensure reproducibility.

In order to improve the overall efficiency and reduce bias of individual models, an ensemble algorithm was developed (MLe-KCNQ2). This MLe-KCNQ2 algorithm consists of the previous LR2, SVC and RF models and predicts on the basis of aggregating the findings of each individual estimator in a soft voting configuration, i.e., on the argmax of the sums of the predicted probabilities [Pedregosa *et al.,* 2011] (Figure 2).

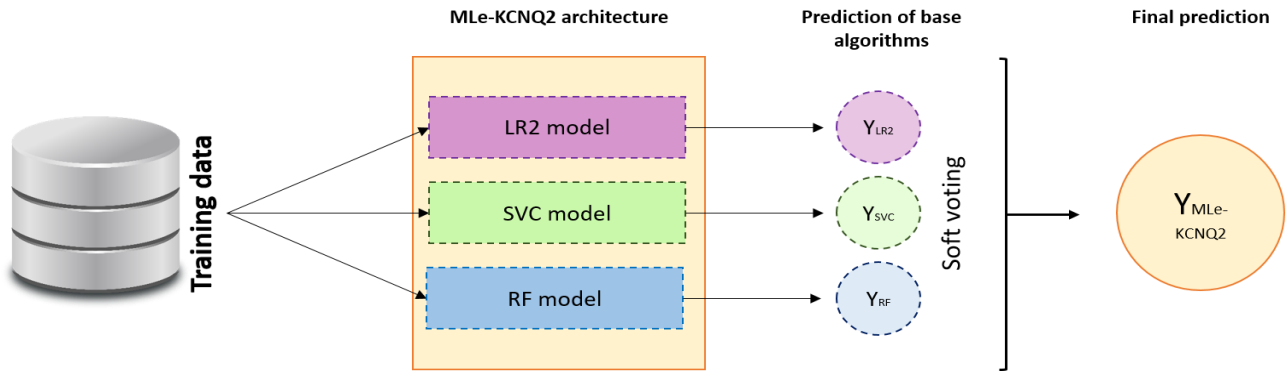

Figure 2. **Architecture of MLe-KCNQ2 algorithm.** MLe-KCNQ2 is an ensemble algorithm consisting of three individual estimators: a Ridge Regression (LR2), a Support Vector Classifier (SVC) and a Random Forest (RF). The final prediction of the MLe-KCNQ2 algorithm was obtained by combining the individual predictions of the base estimators in a soft voting configuration.

MODEL EVALUATION

Models were evaluated by estimating sensitivity and specificity values from confusion matrices. Areas under the Receiver Operating Characteristic (ROC) curves or AUC-ROC were also analyzed. All of these metrics are useful in class imbalance scenarios and are clinically relevant [Japkowicz, 2013]. On the one hand, sensitivity is the metric that evaluates the algorithm's ability to correctly predict true positives (see Equation 1). In the particular case of this study, it would be the ability of the model to distinguish pathogenic variants. On the other hand, specificity is the metric that assesses the ability of the algorithm to correctly differentiate true negatives (see Equation 2). In this context, the ability of the model to distinguish benign variants.

$$Sensitivity = \frac{TP}{TP + FN} \cdot 100 \qquad (1)$$

$$Specificity = \frac{TN}{TN + FP} \cdot 100 \qquad (2)$$

where TP, FP, TN, FN are the true positives, false positives, true negatives and false negatives, respectively.

The ROC curve expresses the relationship between sensitivity and (1 – specificity) for different classification thresholds. The AUC-ROC would be the numerical value that summarizes the predictive ability of that model and is obtained by integrating the area under the ROC curve [Muschelli, 2019]. The AUC-ROC can take any value between 0 and 1 and it is a good indicator of the goodness of the model [Kim *et al.,* 2017]. In general, an AUC-ROC of 0.5 suggests that the model has no discrimination (i.e., would not distinguish

between benign and pathogenic variants), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and above 0.9 is considered outstanding [Hosmer, 2013].

COMPARISON TO ALTERNATIVE APPROACHES

The best model was then compared to some commonly used online predictive tools such as FATHMM (http://fathmm.biocompute.org.uk/disease.html), CONDEL (http://bbglab.irbbarcelona.org/fannsdb/query/condel), PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/bgi.shtml), MutPred2 (http://mutpred.mutdb.org/#qform), SIFT (https://sift.bii.a-star.edu.sg/) and PROVEAN (http://provean.jcvi.org/index.php). All collected variants were used in this comparative analysis.

MODEL PREDICTION

Finally, the model was used to predict the pathogenicity of 293 uncertain significance or conflicting interpretation variants annotated in ClinVar that were not part of the initial dataset. As 3 of these variants were included in both subclasses, only one of them was taken into account. As a result, 290 variants were predicted. In addition to predicting the clinical label, the confidence of each prediction was estimated. A subset of predicted variants was then selected for a biological interpretation, based on the prediction confidence and the availability of electrophysiological studies already reported in the literature.

RESULTS

MACHINE LEARNING MODELS PERFORMANCE

Among individual classifiers, the SVC model had the highest sensitivity, classifying 90.91% of the pathogenic variants. In contrast, the LR2 algorithm was the one with the highest specificity, classifying 90.21% of the benign variants. The RF algorithm resulted in lower sensitivity and specificity than the two previous models, but was also able to correctly discriminate benign and pathogenic variants. The MLe-KCNQ2 ensemble algorithm improved the overall results of individual models as expected, by correctly classifying 91.11% and 90.21% of pathogenic and benign variants, respectively. AUC-ROC values, above 0.9 in most cases, also suggest that the trained models discriminate excellently between benign and pathogenic variants (Table 1). The values shown in Table 1 represent the average of the results obtained for the train and test sets.

Models	Sensitivity (%)	Specificity (%)	AUC-ROC
LR2	90.67	90.21	0.91
SVC	90.91	88.00	0.90
RF	87.87	87.56	0.88
MLe-KCNQ2	91.11	90.21	0.91

Table 1. Performance of trained models. The MLe-KCNQ2 ensemble model performed better, in terms of sensitivity and specificity, than individual Ridge Regression (LR2), Support Vector Classifier (SVC) and Random Forest (RF) classifiers. AUC-ROC values reflect the quality of all models.

RELEVANT FEATURES IN PATHOGENICITY PREDICTION

To determine which variant attributes had the highest influence on pathogenicity prediction; we examined the subset of relevant features previously selected by the feature selection method. This procedure cannot be easily estimated in the ensemble model, as the feature selection is performed in a different way for each individual algorithm of which it is composed. However, an overview of the relevant features in the ensemble model can be obtained by combining the relevant and shared features between the LR2, SVC and RF models

The AlphaFold pLDDT score, the location of the mutation in the FRAG2 of the structural landscape of the K_v7.2 channel (i.e., residues 195-367) and the MTR score were the top 3 relevant features shared by LR2, SVC and RF models. These attributes together with changes in aromaticity (from non-aromatic to non-aromatic, from aromatic to non-aromatic and from non-aromatic to aromatic), the impact of variants on functional domains such as the voltage-sensing domain or topological domains such as the Cytoplasmic or S4 segments, and the conservation across evolution value across species of the substituted residue constitute the top 10 relevant features for predicting the pathogenicity of variants in K_V7.2. Taking into account the three subgroups into which the 16 designed features can be classified (See Methods), 2 out of 3 readily available structural information attributes constituted the top 2 relevant features in pathogenicity prediction of missense variants in *KCNQ2*.

COMPARING MLe-KCNQ2 WITH OTHER PREDICTIVE SOFTWARES

We then compared our novel model with the ability of six *in silico* tools to predict the class of the collected *KCNQ2* missense variants. Performances of these tools are shown in Figure 3. FATHMM showed the top score for sensitivity, classifying 99.67% of the collected pathogenic variants. However, the performance of FATHMM was really poor in terms of specificity as it was not able to correctly classify more

than 8% of the collected benign variants. Similar results were obtained for the rest of the online software. While pathogenic variants were correctly predicted in CONDEL (99.01%), PolyPhen2 (93.44%), MutPred2 (92.46%), SIFT (83.28%) and PROVEAN (92.79%); none of them was able to identify more than 50% of benign variants. Of note, 50% of the benign variants were incorrectly classified by all these tools. These results clearly indicate that the chosen online predictive software did not correctly discriminate between pathogenic and benign variants in *KCNQ2*. In contrast, the novel MLe-KCNQ2 algorithm was able to effectively predict the pathogenicity of *KCNQ2* variants, by correctly classifying both benign and pathogenic variants with sensitivity and specificity scores above 90.2%.

The N780T variant has a large prevalence in the human population

In the course of our analysis, we noted that the N780T variant, which is labeled as benign in ClinVar and LOVD3 databases, was classified as such by all these tools. However, the prevalence of this variant is higher than 60% of the population collected in the gnomAD database, which suggest that it represents allelic variant with little or no pathological consequences.

.

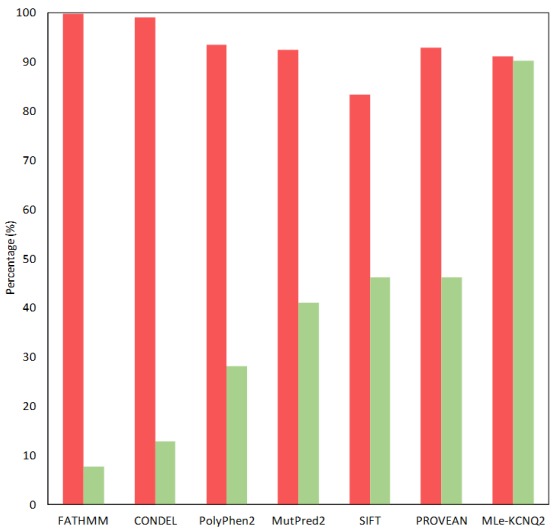

Figure 3. **Performance of MLe-KCNQ2 in comparison to other software.** Sensitivity (red) and specificity (green) scores are compared using the collected data. The novel MLe-KCNQ2 was the best-performing software by being able to identify both benign and pathogenic variants in *KCNQ2* in a similar and efficient manner.

BIOLOGICAL INTERPRETATION OF SELECTED MLe-KCNQ2 PREDICTIONS

Most confidently predicted benign and pathogenic variants by MLe-KCNQ2 are shown in Figure 4. While variants predicted to be pathogenic (Figure 4B and 4C) were mainly found to affect the VSD domain (R144W, W146R, R201C, I205M, T217I and T217A), in particular the S4 segments and the S2-S3 linker, variants affecting the hTW-hB linker were mostly predicted to be benign (R375C, P379L, G380R, P420L, P420R, P420S, P420Q, A421V and P424S) (Figure 4A and 4C). Variants predicted to be pathogenic were also found in the CaM-interaction domain (D355Y, W360R and Y363C), particularly affecting the intracellular hTW, and in the pore domain (W270C). V22M was a variant predicted with "high confidence" to be benign located at the N-terminus of the K_v7.2 channel.

A

Variant	Predicted label	Confidence
P420L	Benign	Very high
P379L	Benign	Very high
P420R	Benign	Very high
P420S	Benign	Very high
P424S	Benign	Very high
A421V	Benign	Very high
P420Q	Benign	Very high
G380R	Benign	Very high
R375C	Benign	Very high
V22M	Benign	Very high

B

Variant	Predicted label	Confidence
W360R	Pathogenic	High
T217I	Pathogenic	High
D355Y	Pathogenic	High
I205M	Pathogenic	High
R201C	Pathogenic	High
R144W	Pathogenic	High
T217A	Pathogenic	High
Y363C	Pathogenic	High
W270C	Pathogenic	High
W146R	Pathogenic	High

C

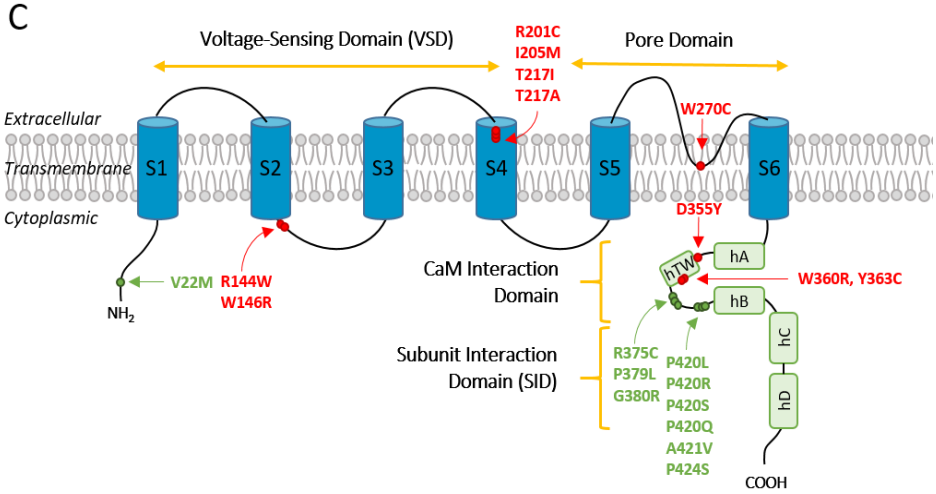

Figure 4. ***KCNQ2* variants for biological interpretation, their pathogenicity prediction and mapping. (A)** Most confidently predicted 10 benign and **(B)** top 10 pathogenic predicted variants using MLe-KCNQ2 software. Benign variants have been predicted with a "very high" confidence value (i.e., with a value equal to or greater than 90%); while pathogenic variants have been predicted with a "high" confidence value (i.e., in the range of equal to or greater than 80% and less than 90%). **(C)** Mapping of these variants on the structure of a monomer of the $K_v7.2$ channel. The picture depicts the main topological and functional domains of a channel monomer. The six transmembrane segments (S1-S6) are shown in blue and the five intracellular helices in green. The functional voltage-sensor domain (VSD) is composed of segments S1-S4, the pore domain of segments S5-S6, the subunit interaction domain (SID) of helices C-D and the calmodulin interaction domain (CaM) of helices hA, hTW and hB. The location of the benign (green) and pathogenic (red) variants is also depicted.

DISCUSSION

Accurate pathogenicity prediction of missense variants is of great importance in genetic studies and clinical diagnosis [Qi et al., 2021]. In clinical practice, when a genetic variant is found in the setting of disease, the ACMG (American College of Medical Genetics and Genomics) recommends a variety of techniques to be employed in an attempt to classify the variant as either benign or pathogenic [Richards *et al.,* 2015]. However, since the production of genome sequencing data has become routine, the interpretation of the large number of *de novo* discovered variants is posing a major challenge [Lappalainen *et al.,* 2019]. *In silico* prediction softwares go some way to address this major issue, but often yield inconsistent results [Chun & Fay, 2009]. ML techniques offer an opportunity to improve pathogenicity prediction tools [Draelos *et al.,* 2022].

The structure of a protein is intimately linked to its stability, function and interactions [Ittisoponpisan *et al.*, 2019]. It is therefore expected that the structure would be capable of answering a vast number of biological questions including the pathogenicity of variants [Burley *et al.,* 2018; Ginalski, 2016]. Despite efforts over many years, only 35% of human proteins map to a PDB entry with the 3D atomic coordinates, and in many cases the structure covers only some fragments of the sequence [Tunyasuvunakool *et al.,* 2021; Bateman *et al.,* 2021]. For instance, in the structure obtained for $K_v7.2$ by cryo-EM recently, the spatial location of 523 amino acids (~ 60% of the entire sequence) are not present in the PDB file.

This work pursued two main objectives: (1) developing a ML model to improve accuracy for predicting the pathogenicity of *KCNQ2* missense variants by combining biological-evolutive information, physico-chemical changes and structural features; and (2) creating a framework that can be extended to the prediction of the pathogenicity of variants occurring in other genes. Although there was currently no specific ML based software for *KCNQ2*, similar approaches have been applied to *KCNQ1*, another member of the *KCNQ* family [Phul *et al.,* 2022; Draelos *et al.,* 2022; Li *et al.,* 2017].

First of all, we characterized a set of collected variants combining traditional evolutive features such as evolutive conservation value of the substituted residue; biological information such as mutated residue or affected $K_v7.2$ channel domains and physico-chemical changes (i.e., charge, hydrophobicity, molecular weight…); with easily-obtainable genomic and structural information. The incorporation of MTR score and AlphaFold2 pLDDT value constitutes an innovative feature about MLe-KCNQ2 software, which sets it apart from other tools. As a result, a combination of biological-evolutive information, physico-chemical changes and structural features have been considered for pathogenicity prediction. Among the features that influenced the classification of *KCNQ2* variants by MLe-KCNQ2, designed structural descriptors were the most relevant as we expected, followed by the MTR score. The addition of these three novel features in MLe-KCNQ2 allowed increasing the predictive power, resulting in sensitivity and specificity scores of 91.11% and 90.21%, respectively.

We also compared the prediction performances of six *in silico* tools to that of MLe-KCNQ2, and found that the presented gene-specific model outperformed the other softwares. None of the online tools were able to identify benign variants in *KCNQ2* with adequate values (<8-50% in all cases), although pathogenic variants were correctly classified by all of them. Our results suggested that commonly used *in silico* tools could not correctly discriminate between pathogenic and benign variants in *KCNQ2*. We believe that these results can be explained by the type of variants with which each of the algorithms has been trained. While tools like FATHMM, CONDEL, PolyPhen2, MutPred2, SIFT or PROVEAN were trained using information from genome-wide genetic variation data [Pejaver *et al.,* 2020; Shihab *et al.,* 2013; Choi *et al.,* 2012; González-Pérez & López-Bigas, 2011; Adzhubei *et al.,* 2010; Ng & Henikoff, 2001]; MLe-KCNQ2 was specifically trained with *KCNQ2* missense variants. This reflects that development of genome-wide tools was based on heterogeneous datasets including a wide range of proteins with diverse functions and associated diseases [Phul *et al.,* 2022]. As a direct consequence, their prediction accuracy can vary between genes [Richards *et al.,* 2015]. In contrast, MLe-KCNQ2 was trained with specific variants of the *KCNQ2* gene with an emphasis on its particularities, such as the class imbalance scenario between benign and pathogenic

variants as well as the lack of experimental structure for the design of 3D-features. From the outset, our software was designed to fit the *KCNQ2* gene problem. Therefore, it makes sense that MLe-KCNQ2 had better classification metrics than these genome-wide tools.

As the results of MLe-KCNQ2 were promising compared to commonly employed tools we decided to give a biological interpretation of a subset of selected variants in order to justify MLe-KCNQ2 predictions and highlight their usefulness. As electrophysiological studies were not available for all the variants, we obtained this information from some experimental studies recently reported in the literature. These works allowed us to justify 3 variants. Firstly, T271I mutation is located in the S4 domain of the $K_V7.2$ channel and was predicted as a pathogenic variant. The S4 domain is part of the voltage sensor, has 4 positively charged arginine residues and its movement in response to changes in membrane potential is the basis for the voltage sensitivity of these channels [Sun & MacKinnon, 2017; Cui, 2016; Nakajo & Kubo, 2015]. In addition, experimental studies have shown that mutations in the S4 domain cause significant changes in the activation of $K_v7.2$ and would be associated with a pathogenic clinical picture [Zhang *et al.,* 2020; Orhan *et al.,* 2014]. Recent studies done at the Fudan University Children's Hospital in China in infants who had seizures in the first month of life determine that this mutation results in encephalopathy [Xu *et al.,* 2021]. Secondly, the R144W mutation is located in the S2 transmembrane segment and is therefore part of the voltage-sensing domain. It was also classified as a pathogenic variant by MLe-KCNQ2. Functional studies performed by whole-cell patch-clamp show that activation of homomeric $K_v7.2$ R144W/Q/G channels is shifted to the left, suggesting gain-of-function effects (i.e., the probability of opening the gate increases at depolarized potentials) and a pathogenic clinic [Miceli *et al.,* 2022]. Finally, the W360R mutation is located in the C-terminal region of the $K_v7.2$ channel, specifically in the hTW. MLe-KCNQ2 also predicted W360R as a pathogenic variant with high confidence. Helix TW is located between helices A and B, within the calmodulin (CaM) binding domain of the protein. In addition, the interaction with CaM could be affected in the three-dimensional space due to the substitution of tryptophan (W), non-polar amino acid, by arginine (R), a polar residue. Structurally, this change in polarity could lead to a conformational effect sufficiently relevant for the interaction with CaM to be affected in T359A and, therefore, lead to a pathological clinical scenario. Furthermore, experimental studies have demonstrated the importance of residues extending from T359 to Y362 for CaM binding and it has been suggested that the reduction of this interaction could explain their pathogenicity [Gomis-Perez *et al.,* 2015; Richards *et al.,* 2004]. In all cases, the predictions made for T217I, R144W and W360R by MLe-KCNQ2 are supported by previous studies.

Finally, in this work we created a framework to develop AI algorithms that can be extended to the prediction of the pathogenicity of variants occurring in other genes. We also demonstrated the extrapolation capability of the working scheme by creating a preliminary model for *KCNH2* gene. The sensitivity and specificity values achieved, higher than 90%, demonstrate that the scheme followed for MLe-KCNQ2 software can be the starting point for the development of other gene-specific predictive software.

STUDY LIMITATIONS AND FUTURE DIRECTIONS

This work has, among others, four main limitations. The first is that only *KCNQ2* missense variants were included in the design of the MLe-KCNQ2 algorithm. As a direct consequence, MLe-KCNQ2 may not accurately predict neither other *KCNQ2* molecular variants (e.g., frameshift, nonsense…) nor missense mutations in other voltage-gated potassium channels. The second limitation is that MLe-KCNQ2 cannot differentiate between the wide range of phenotypes in the *KCNQ2* encephalopathies. It also cannot determine whether a mutation predicted as pathogenic involves a gain-of-function or a loss-of-function. Pathogenic variants may have a range of effect size and penetrance [Qi et al., 2021]. These difficulties make it necessary to carry out complementary experiments that validate the ML result will help to build trust in approaches and outputs [Vamathevan *et al,* 2019]. This is the main reason why a molecular validation of the predicted variants is highly recommended. The third limitation is the size of the dataset. Although a substantial effort was spent in the compilation of unique *KCNQ2* missense variants, only 353 had a clear or

sufficiently reliably clinic description. When larger data sets become available, better predictions could be achieved with MLe-KCNQ2. The last limitation would be the variant reclassifications that regularly occur in clinical databases such as Clinvar [McInnes *et al.,* 2021]. If the training data is not updated as these reclassifications occur, any ML-based model would be obsolete and would diminish its clinical relevance.

With available data becoming "larger", ML algorithms are going to systematically generate improved outputs and new interesting applications are expected to follow [Vamathevan *et al,* 2019]. As a consequence, we suggest further work on the dataset for missense variants of *KCNQ2* by incorporating unseen variants from gnomAD database (https://gnomad.broadinstitute.org/) or recently reported studies. Designing new features for variant characterization such as change in number of hydrogen donor sites or change in number of hydrogen acceptor sites would improve classification metrics as some works remark [Phul *et al.,* 2022]. Advances in protein structure prediction (e.g., AlphaFold2) as well as cryo-EM technologies could lead in the design of more complex 3D-features that could make a breakthrough in the prediction of variant pathogenicity. More complex artificial intelligence (AI) methods can also be employed, for example, deep learning techniques. Although there are already a number of studies that have worked with neural networks, one major drawback of deep learning is that it requires an immense amount of data [Jakhar & Kaur, 2019]. Transfer learning offers an opportunity to leverage the power of deep learning in situations where data is limited [McInnes *et al.,* 2021]. This emerging approach consists of training a model using data from a well-studied gene (X) and then refine the model with data from a poorly studied gene (Y). The resulting model may perform very well on Y because the ''lessons'' learned in modeling X transfer well to Y [Taroni *et al.,* 2019].

Despite these limitations, MLe-KCNQ2 software represents a major advance in the clinical diagnosis of *KCNQ2*-related encephalopathies. This work has provided new opportunities in the study of the $K_v7.2$ channel that are expected to be further exploited in the future. In addition, the analysis of the pathogenic landscape could help identifying previously unknown domains within the protein sequence, and to discover new functions and possible new pathways for treatments.

CONCLUSIONS

Here, we provide a powerful tool to predict the pathogenicity of KCNQ2 missense variants. The strength of MLe-KCNQ2 software relies on its performance with both types of variants, benign and pathogenic, with an estimated specificity and sensitivity above 90%. We also found that combining traditional physico-chemical and biological-evolutive features with easily-obtainable structural information could be the key to creating an optimal predictive model. MLe-KCNQ2 will be a helpful tool to improve the accuracy of genetic diagnoses for *KCNQ2*-related encephalopathies. The framework employed here can be extended to the prediction of the pathogenicity of variants occurring in other genes, as we demonstrated with hERG ion channel.

SUPPLEMENTARY MATERIAL

Supplementary Material, code, and original data are available at https://github.com/KCNQlab

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., . . . Teodoro, D. (2020). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), D480-D489. https://doi.org/10.1093/nar/gkaa1100

Biervert, C., Schroeder, B. C., Kubisch, C., Berkovic, S. F., Propping, P., Jentsch, T. J., & Steinlein, O. K. (1998). A Potassium Channel Mutation in Neonatal Human Epilepsy. Science, 279(5349), 403–406. https://doi.org/10.1126/science.279.5349.403

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L. D., Christie, C., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Liang, Y., Lowe, R., Peisach, E., . . . Ioannidis, Y. E. (2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research, 47(D1), D520-D528. https://doi.org/10.1093/nar/gky949

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE, 7(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. Genome Research, 19(9), 1553–1561. https://doi.org/10.1101/gr.092619.109

Cooper, G. M., & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics, 12(9), 628–640. https://doi.org/10.1038/nrg3046

Cui, J. (2016). Voltage-Dependent Gating: Novel Insights from KCNQ1 Channels. Biophysical Journal, 110(1), 14–25. https://doi.org/10.1016/j.bpj.2015.11.023

David, A., Islam, S., Tankhilevich, E. & Sternberg, M. J. (2022). The AlphaFold Database of Protein Structures: A Biologist's Guide. Journal of Molecular Biology, 434(2), 167336. https://doi.org/10.1016/j.jmb.2021.167336

DeLano WL. Pymol: An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography. 2002.

Devaux, J. J. (2004). KCNQ2 Is a Nodal K+ Channel. Journal of Neuroscience, 24(5), 1236–1244. https://doi.org/10.1523/jneurosci.4512-03.2004

Draelos, R. L., Ezekian, J. E., Zhuang, F., Moya-Mendez, M. E., Zhang, Z., Rosamilia, M. B., Manivannan, P. K., Henao, R., & Landstrom, A. P. (2022). GENESIS: Gene-Specific Machine Learning Models for Variants of Uncertain Significance Found in Catecholaminergic Polymorphic Ventricular Tachycardia and Long QT Syndrome-Associated Genes. Circulation: Arrhythmia and Electrophysiology, 15(4). https://doi.org/10.1161/circep.121.010326

Etxeberria, A., Aivar, P., Rodriguez-Alfaro, J. A., Alaimo, A., Villace, P., Gomez-Posada, J. C., Areso, P., & Villarroel, A. (2007). Calmodulin regulates the trafficking of KCNQ2 potassium channels. The FASEB Journal, 22(4), 1135–1143. https://doi.org/10.1096/fj.07-9712com

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., & Bateman, A. (2009). The Pfam protein families database. Nucleic Acids Research, 38, D211-D222. https://doi.org/10.1093/nar/gkp985

Ginalski, K. (2006). Comparative modeling for protein structure prediction. Current Opinion in Structural Biology, 16(2), 172–177. https://doi.org/10.1016/j.sbi.2006.02.003

Gomis-Perez, C., Alaimo, A., Fernandez-Orth, J., Alberdi, A., Aivar-Mateo, P., Bernardo-Seisdedos, G., Malo, C., Areso, P., Felipe, A., & Villarroel, A. (2015). Unconventional calmodulin anchoring site within the AB module of Kv7.2 channels. Journal of Cell Science. https://doi.org/10.1242/jcs.174128

González-Pérez, A., & López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. The American Journal of Human Genetics, 88(4), 440–449. https://doi.org/10.1016/j.ajhg.2011.03.004

Goto, A., Ishii, A., Shibata, M., Ihara, Y., Cooper, E. C., & Hirose, S. (2019). Characteristics of KCNQ2 variants causing either benign neonatal epilepsy or developmental and epileptic encephalopathy. Epilepsia, 60(9), 1870–1880. https://doi.org/10.1111/epi.16314

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology, 23(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Gutman, G. A., Chandy, K. G., Adelman, J. P., Aiyar, J., Bayliss, D. A., Clapham, D. E., Covarriubias, M., Desir, G. V., Furuichi, K., Ganetzky, B., Garcia, M. L., Grissmer, S., Jan, L. Y., Karschin, A., Kim, D., Kuperschmidt, S., Kurachi, Y., Lazdunski, M., Lesage, F., . . . Wymore, R. S. (2003). International Union of Pharmacology. XLI. Compendium of Voltage-Gated Ion Channels: Potassium Channels. Pharmacological Reviews, 55(4), 583–586. https://doi.org/10.1124/pr.55.4.9

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hoque, N., Bhattacharyya, D., & Kalita, J. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371–6385. https://doi.org/10.1016/j.eswa.2014.04.019

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.

Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: Visual molecular dynamics. Journal of Molecular Graphics, 14(1), 33-38. https://doi.org/10.1016/0263-7855(96)00018-5

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. (2019). Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? Journal of Molecular Biology, 431(11), 2197–2212. https://doi.org/10.1016/j.jmb.2019.04.009

Jakhar, D., & Kaur, I. (2019). Artificial intelligence, machine learning and deep learning: definitions and differences. Clinical and Experimental Dermatology, 45(1), 131–132. https://doi.org/10.1111/ced.14029

Japkowicz, N. (2013). Assessment Metrics for Imbalanced Learning. Imbalanced Learning, 187–206. https://doi.org/10.1002/9781118646106.ch8

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., ŽÍdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kato, M., Yamagata, T., Kubota, M., Arai, H., Yamashita, S., Nakagawa, T., FujII, T., Sugai, K., Imai, K., Uster, T., Chitayat, D., Weiss, S., Kashii, H., Kusano, R., Matsumoto, A., Nakamura, K., Oyazato, Y., Maeno, M., Nishiyama, K., . . . Saitsu, H. (2013). Clinical spectrum of early onset epileptic encephalopathies caused byKCNQ2mutation. Epilepsia, 54(7), 1282–1287. https://doi.org/10.1111/epi.12200

Katsonis, P., Wilhelm, K., Williams, A., & Lichtarge, O. (2022). Genome interpretation using in silico predictors of variant impact. Human Genetics. https://doi.org/10.1007/s00439-022-02457-6

Kim, H. J., Yang, D., Kim, S. H., Won, D., Kim, H. D., Lee, J. S., Choi, J. R., Lee, S. T., & Kang, H. C. (2021b). Clinical characteristics of KCNQ2 encephalopathy. Brain and Development, 43(2), 244–250. https://doi.org/10.1016/j.braindev.2020.08.015

Kim, S. J., Cho, K. J., & Oh, S. (2017). Development of machine learning models for diagnosis of glaucoma. PLOS ONE, 12(5), e0177726. https://doi.org/10.1371/journal.pone.0177726

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. Cell, 177(1), 70–84. https://doi.org/10.1016/j.cell.2019.02.032

Larrea-Sebal, A., Benito-Vicente, A., Fernandez-Higuero, J. A., Jebari-Benslaiman, S., Galicia-Garcia, U., Uribe, K. B., Cenarro, A., Ostolaza, H., Civeira, F., Arrasate, S., González-Díaz, H., & Martín, C. (2021). MLb-LDLr. JACC: Basic to Translational Science, 6(11), 815–827. https://doi.org/10.1016/j.jacbts.2021.08.009

Leong, I. U., Stuckey, A., Lai, D., Skinner, J. R., & Love, D. R. (2015). Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. BMC Medical Genetics, 16(1). https://doi.org/10.1186/s12881-015-0176-z

Li, B., Mendenhall, J. L., Kroncke, B. M., Taylor, K. C., Huang, H., Smith, D. K., Vanoye, C. G., Blume, J. D., George, A. L., Sanders, C. R., & Meiler, J. (2017). Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance. Circulation: Cardiovascular Genetics, 10(5). https://doi.org/10.1161/circgenetics.117.001754

MacArthur, D. G., & Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Human Molecular Genetics, 19(R2), R125-R130. https://doi.org/10.1093/hmg/ddq365

Marrion, N. V. (1997). CONTROL OF M-CURRENT. Annual Review of Physiology, 59(1), 483–504. https://doi.org/10.1146/annurev.physiol.59.1.483

Mayer, A. N., Dimmock, D. P., Arca, M. J., Bick, D. P., Verbsky, J. W., Worthey, E. A., Jacob, H. J., & Margolis, D. A. (2011). A timely arrival for genomic medicine. Genetics in Medicine, 13(3), 195–196. https://doi.org/10.1097/gim.0b013e3182095089

McInnes, G., Sharo, A. G., Koleske, M. L., Brown, J. E., Norstad, M., Adhikari, A. N., Wang, S., Brenner, S. E., Halpern, J., Koenig, B. A., Magnus, D. C., Gallagher, R. C., Giacomini, K. M., & Altman, R. B. (2021). Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. The American Journal of Human Genetics, 108(4), 535–548. https://doi.org/10.1016/j.ajhg.2021.03.003

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-00a

Miceli, F., Millevert, C., Soldovieri, M. V., Mosca, I., Ambrosino, P., Carotenuto, L., Schrader, D., Lee, H. K., Riviello, J., Hong, W., Risen, S., Emrick, L., Amin, H., Ville, D., Edery, P., de Bellescize, J., Michaud, V., Van-Gils, J., Goizet, C., . . . Weckhuysen, S. (2022b). KCNQ2 R144 variants cause neurodevelopmental disability with language impairment and autistic features without neonatal seizures through a gain-of-function mechanism. eBioMedicine, 81, 104130. https://doi.org/10.1016/j.ebiom.2022.104130

Millichap, J. J., Park, K. L., Tsuchida, T., Ben-Zeev, B., Carmant, L., Flamini, R., Joshi, N., Levisohn, P. M., Marsh, E., Nangia, S., Narayanan, V., Ortiz-Gonzalez, X. R., Patterson, M. C., Pearl, P. L., Porter, B., Ramsey, K., McGinnis, E. L., Taglialatela, M., Tracy, M., . . . Cooper, E. C. (2016). KCNQ2encephalopathy. Neurology Genetics, 2(5), e96. https://doi.org/10.1212/nxg.0000000000000096

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS). https://doi.org/10.1109/icics49469.2020.239556

Muschelli, J. (2019). ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. Journal of Classification, 37(3), 696–708. https://doi.org/10.1007/s00357-019-09345-1

Nakajo, K., & Kubo, Y. (2015). KCNQ1 channel modulation by KCNE proteins via the voltage-sensing domain. The Journal of Physiology, 593(12), 2617–2625. https://doi.org/10.1113/jphysiol.2014.287672

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Research, 11(5), 863–874. https://doi.org/10.1101/gr.176601

Numis, A. L., Angriman, M., Sullivan, J. E., Lewis, A. J., Striano, P., Nabbout, R., & Cilio, M. R. (2013). KCNQ2 encephalopathy: Delineation of the electroclinical phenotype and treatment response. Neurology, 82(4), 368–370. https://doi.org/10.1212/wnl.0000000000000060

Orhan, G., Bock, M., Schepers, D., Ilina, E. I., Reichel, S. N., Löffler, H., Jezutkovic, N., Weckhuysen, S., Mandelstam, S., Suls, A., Danker, T., Guenther, E., Scheffer, I. E., de Jonghe, P., Lerche, H., & Maljevic, S. (2014). Dominant-negative effects ofKCNQ2mutations are associated with epileptic encephalopathy. Annals of Neurology, 75(3), 382–394. https://doi.org/10.1002/ana.24080

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-19669-x

Phul, S., Kuenze, G., Vanoye, C. G., Sanders, C. R., George, A. L., & Meiler, J. (2022). Predicting the functional impact of KCNQ1 variants with artificial neural networks. PLOS Computational Biology, 18(4), e1010038. https://doi.org/10.1371/journal.pcbi.1010038

Pisano, T., Numis, A. L., Heavin, S. B., Weckhuysen, S., Angriman, M., Suls, A., Podesta, B., Thibert, R. L., Shapiro, K. A., Guerrini, R., Scheffer, I. E., Marini, C., & Cilio, M. R. (2015). Early and effective treatment of KCNQ2 encephalopathy. Epilepsia, 56(5), 685–691. https://doi.org/10.1111/epi.12984

Polat, K., & Güneş, S. (2009). A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Systems with Applications, 36(7), 10367–10373. https://doi.org/10.1016/j.eswa.2009.01.041

Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J. J., Chung, W. K., Guan, Y., & Shen, Y. (2021). MVP predicts the pathogenicity of missense variants by deep learning. Nature Communications, 12(1). https://doi.org/10.1038/s41467-020-20847-0

Richards, M. C. (2004). Novel mutations in the KCNQ2 gene link epilepsy to a dysfunction of the KCNQ2-calmodulin interaction. Journal of Medical Genetics, 41(3), 35e-335. https://doi.org/10.1136/jmg.2003.013938

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). 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. Genetics in Medicine, 17(5), 405–424. https://doi.org/10.1038/gim.2015.30

Sallah, S. R., Sergouniotis, P. I., Barton, S., Ramsden, S., Taylor, R. L., Safadi, A., Kabir, M., Ellingford, J. M., Lench, N., Lovell, S. C., & Black, G. C. M. (2020). Using an integrative machine learning approach utilising homology modelling to clinically interpret genetic variants: CACNA1F as an exemplar. European Journal of Human Genetics, 28(9), 1274–1282. https://doi.org/10.1038/s41431-020-0623-y

Schwake, M. (2006). Structural Determinants of M-Type KCNQ (Kv7) K+ Channel Assembly. Journal of Neuroscience, 26(14), 3757–3766. https://doi.org/10.1523/jneurosci.5017-05.2006

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics, 29(12), 1504–1510. https://doi.org/10.1093/bioinformatics/btt182

Singh, N. A., Charlier, C., Stauffer, D., DuPont, B. R., Leach, R. J., Melis, R., Ronen, G. M., Bjerre, I., Quattlebaum, T., Murphy, J. V., McHarg, M. L., Gagnon, D., Rosales, T. O., Peiffer, A., Anderson, V. E., & Leppert, M. (1998). A novel potassium channel gene, KCNQ2, is mutated in an inherited epilepsy of newborns. Nature Genetics, 18(1), 25–29. https://doi.org/10.1038/ng0198-25

Sun, J., & MacKinnon, R. (2017). Cryo-EM Structure of a KCNQ1/CaM Complex Reveals Insights into Congenital Long QT Syndrome. Cell, 169(6), 1042–1050.e9. https://doi.org/10.1016/j.cell.2017.05.019

Taroni, J. N., Grayson, P. C., Hu, Q., Eddy, S., Kretzler, M., Merkel, P. A., & Greene, C. S. (2019). MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Systems, 8(5), 380–394.e4. https://doi.org/10.1016/j.cels.2019.04.003

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation, 32(4), 358–368. https://doi.org/10.1002/humu.21445

Traynelis, J., Silk, M., Wang, Q., Berkovic, S. F., Liu, L., Ascher, D. B., Balding, D. J., & Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Research, 27(10), 1715–1729. https://doi.org/10.1101/gr.226589.117

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., . . . Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. Nature, 596(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

Wainberg, M., Merico, D., Delong, A., & Frey, B. J. (2018). Deep learning in biomedicine. Nature Biotechnology, 36(9), 829–838. https://doi.org/10.1038/nbt.4233

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A., Rempfer, C., Bordoli, L., Lepore, R. & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research, 46(W1), W296-W303. https://doi.org/10.1093/nar/gky427

Weckhuysen, S., Ivanovic, V., Hendrickx, R., van Coster, R., Hjalgrim, H., Moller, R. S., Gronborg, S., Schoonjans, A. S., Ceulemans, B., Heavin, S. B., Eltze, C., Horvath, R., Casara, G., Pisano, T., Giordano, L., Rostasy, K., Haberlandt, E., Albrecht, B., Bevot, A., . . . de Jonghe, P. (2013). Extending the KCNQ2 encephalopathy spectrum: Clinical and neuroimaging findings in 17 patients. Neurology, 81(19), 1697–1703. https://doi.org/10.1212/01.wnl.0000435296.72400.a1

Weckhuysen, S., Mandelstam, S., Suls, A., Audenaert, D., Deconinck, T., Claes, L. R., Deprez, L., Smets, K., Hristova, D., Yordanova, I., Jordanova, A., Ceulemans, B., Jansen, A., Hasaerts, D., Roelens, F., Lagae, L., Yendle, S., Stanley, T., Heron, S. E., . . . Peter De Jonghe. (2012). KCNQ2 encephalopathy: Emerging phenotype of a neonatal epileptic encephalopathy. Annals of Neurology, 71(1), 15–25. https://doi.org/10.1002/ana.22644

Xenakis, M. N., Kapetis, D., Yang, Y., Gerrits, M. M., Heijman, J., Waxman, S. G., Lauria, G., Faber, C. G., Westra, R. L., Lindsey, P. J., & Smeets, H. J. (2021). Hydropathicity-based prediction of pain-causing NaV1.7 variants. BMC Bioinformatics, 22(1). https://doi.org/10.1186/s12859-021-04119-2

Xu, Y., Dou, Y. L., Chen, X., Dong, X. R., Wang, X. H., Wu, B. B., Cheng, G. Q., & Zhou, Y. F. (2021). Early initial video-electro-encephalography combined with variant location predict prognosis of KCNQ2-related disorder. BMC Pediatrics, 21(1). https://doi.org/10.1186/s12887-021-02946-z

Zhang, J., Kim, E. C., Chen, C., Procko, E., Pant, S., Lam, K., Patel, J., Choi, R., Hong, M., Joshi, D., Bolton, E., Tajkhorshid, E., & Chung, H. J. (2020). Identifying mutation hotspots reveals pathogenetic mechanisms of KCNQ2 epileptic encephalopathy. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-61697-6