Access this article online

Quick Response Code:

Website:
www.neurologyindia.com

DOI:
10.4103/neuroindia.
NI_261_21

TREADS: Target Research for Anti-epileptic Drugs Using Data Science

Janaki Chintalapati, Arvind Kumar, Hosur MV[1], Supriya N Pal

Abstract:

Context: Epilepsy is a common neurological disease and is classified into different types based on features such as the kind of seizure, age of onset, part of brain effected, etc. There are nearly 30 approved anti-epileptic drugs (AEDs) for treating different epilepsies and each drug targets proteins exhibiting a specific molecular mechanism of action. There are many genes, proteins, and microRNAs known to be associated with different epileptic disorders. This rich information on epilepsy-associated data is not available at one single location and is scattered across multiple publicly available repositories. There is a need to have a single platform integrated with the data, as well as tools required for epilepsy research.

Methods and Material: Text mining approaches are used to extract data from multiple biological sources. The data is curated and populated within an in-house developed epilepsy database. Machine-learning based models are built in-house to know the probability of a protein being druggable based on the significant protein features. A web interface is provided for the access of the epilepsy database as well as the ML-based tool developed in-house.

Results: The epilepsy-associated data is made accessible through a web browser. For a protein of interest, the platform provides all the feature values, and the results generated using different machine learning models are displayed as visualization plots.

Conclusions: To meet these objectives, we present TREADS, a platform for epilepsy research community, having both database and an ML-based tool for the study of AED targets.

To access TREADS: https://treads-aer.cdacb.in

Key Words:

Anti-epileptic drugs, drug targets, epilepsy-associated genes, epilepsy database, epilepsy syndromes

Key Messages:

TREADS is a computational platform built for the benefit of the epilepsy research community. The platform, deployed as a website, provides access to the in-house developed epilepsy database and the machine learning models trained to identify targets based on their druggable properties.

Epilepsy, the fourth most common neurological disorder in the world, is characterized by recurrent, and usually unprovoked seizures. There are more than 50 million people with epilepsy (PWE) in the world and more than 10 million in India.[1] Nearly 30% of these patients are resistant to the known anti-epileptic drugs (AEDs). There are nearly 30 approved AEDs reported in DrugBank[2] and a few drugs are still in the investigational phase. Each drug targets specific proteins and exhibits specific molecular mechanisms of action. Based on the experimental studies carried out by the epilepsy research community, it is observed that there are hundreds of epilepsy-associated genes[3,4] and microRNAs,[5] and each of these are either specific to a particular epileptic disorder or may be associated with many disorders. This data from experimental or computational studies is only available in the published literature and there is no single database hosting the complete information related to epilepsy. CarpeDB (http://carpedb.ua.edu/), an epilepsy genetic database, has not been updated since 2010 and has many static links. EpilepsyGene, a resource for genes and mutations related to epilepsy,[6] is currently non-functional. Other

C-DAC Knowledge Park, No.1, Old Madras Road, Byapanahalli, [1]Adjunct Faculty, School of Natural Sciences and Engineering, National Institute of Advanced Studies, IISc Campus, Bangalore, Karnataka, India

Address for correspondence:
Dr. Janaki Chintalapati,
C-DAC Knowledge Park,
No.1, Old Madras Road,
Byapanahalli, Bangalore,
Karnataka, India.
E-mail: janaki@cdac.in

How to cite this article: Chintalapati J, Kumar A, Hosur MV, Pal SN. TREADS: Target Research for Anti-epileptic Drugs Using Data Science. Neurol India 2024;72:620-5.

Submitted: 15-Mar-2021 **Revised:** 31-Mar-2021
Accepted: 09-Mar-2022 **Published:** 30-Jun-2024

epilepsy databases majorly host EEG datasets.[7] We felt the need to develop a database for epilepsy that has data integrated from various public domain biological repositories. Having such a resourceful database can come handy to the research community in extracting relevant information. AutDB is one such condition-driven database having all the genes associated to Autism Spectrum Disorder.[8] In this paper, we present TREADS: a platform to access in-house developed epilepsy database as well as machine-learning-based models to identify potential targets.

Subjects and Methods

Building of epilepsy database

The epilepsy-associated genes and microRNAs experimentally studied are extracted from the published literature (PubMed) using text mining tools. From DrugBank, the AEDs and their corresponding targets are extracted using web scraping. For each AED target, the corresponding protein data is pulled from the UniProt XML files. This data is populated to the in-house developed epilepsy database.

For identifying druggable proteins, 20,268 human reviewed proteins are considered as input and the following feature values and machine learning algorithms are considered. The properties studied as features include frequency of each of the 20 amino acids, iso-electric point and hydrophobicity for every protein, the number of low-complexity regions (LCR), PEST motifs (Proline, Glutamic acid, Serine and Threonine signature, i.e., hydrophilic motifs), secondary structure elements (i.e., number of beta strands, alpha helices, and coils). Also, the post-translational modifications which include the number of O-linked and N-linked glycosylation sites, number of phosphoserines, phosphothreonines, and phosphotyrosines are considered. These properties are considered to be relevant from previous studies,[9] and are hence used as features for building the machine learning models. Permutation importance was used for calculating the feature importance employing random forest. The work done on the feature extraction and machine learning models built in-house for predicting druggability of a protein is published in ICMLANT 2020.[10] The 20,268 human reviewed proteins are given as input to these ML-based classifier models; different base classifiers such as logistic regression, support vector machine (SVM), decision tree, and multi-layer perceptron were build using the above-mentioned properties. Random forest, a basic ensemble of learning techniques and two stacking approaches were also tried for the same dataset. The first version of stacking works by adding the predictions from the base learners to the test and train data as new features (retaining initial features) for the meta learner. The second version of stacking follows the same principal as the first, the difference is that it takes only the predictions from the base learners and drops all of the initial features for the meta learner to learn on. The data is split into 80% for training and 20% for testing, based on stratified sampling technique and to ensure there is no bias in the dataset. Accuracy and F1-score, that is, harmonic mean of precision and recall are used as performance metrics. The accuracy using multilayer perceptron, SVM, random forest, and gradient-boosting models was found to be good, that is, more than 70% for the given dataset, and the F1-score was in the range of 0.4 and 0.5 for all of the four models. Hence, the same models have been used in TREADS. In future, the results using optimized stacking approach models would be added. 15.6% of the input dataset, that is, 3,162 of the 20,268 human reviewed proteins were found to have druggable properties using these four models. The results of the machine learning (ML) models are made accessible through the platform. For each protein, external links are provided to UniProt,[11] Bgee gene expression data,[12] DrugBank[2] and Ensembl repositories. A comprehensive database is developed using Postgre SQL and data from these repositories is populated [Figure 1].

Platform development

The web interface from which the database as well as ML models could be accessed is developed using Django framework. Both search and browse options are provided for accessing the data.

Results

TREADS enables a user to browse epilepsy-associated data.

Advantages of the Platform:

1. Search for epilepsy-associated proteins using UniProt or Ensembl identifier or protein name.
2. Browse for genes associated with a particular epileptic disorder or those common across multiple disorders.
3. Browse AEDs, AED targets, Epilepsy-associated genes, microRNAs (miRNAs), and pathways.
4. ML-based models for identifying proteins with druggable properties.
5. Comparative visualization plots for protein properties.
6. For a given protein, display of literature associated with epilepsy studies if available.
7. Data downloadable in tabular format and can be saved in Excel, PDF, or CSV format.
8. Mobile-friendly.

Search option

TREADS can be used to search information of any reviewed human protein regarding druggable features, interactive plots for those features, scientific literature associating that protein with epilepsy, and corresponding anti-epileptic drugs if the searched protein is a known drug target. In case the searched protein is not a known drug target, the user can also see machine learning predictions stating that whether the searched protein can be a probable drug target or not. The user can either enter the gene name, or the UniProt or Ensembl identifier for any protein to search for details.

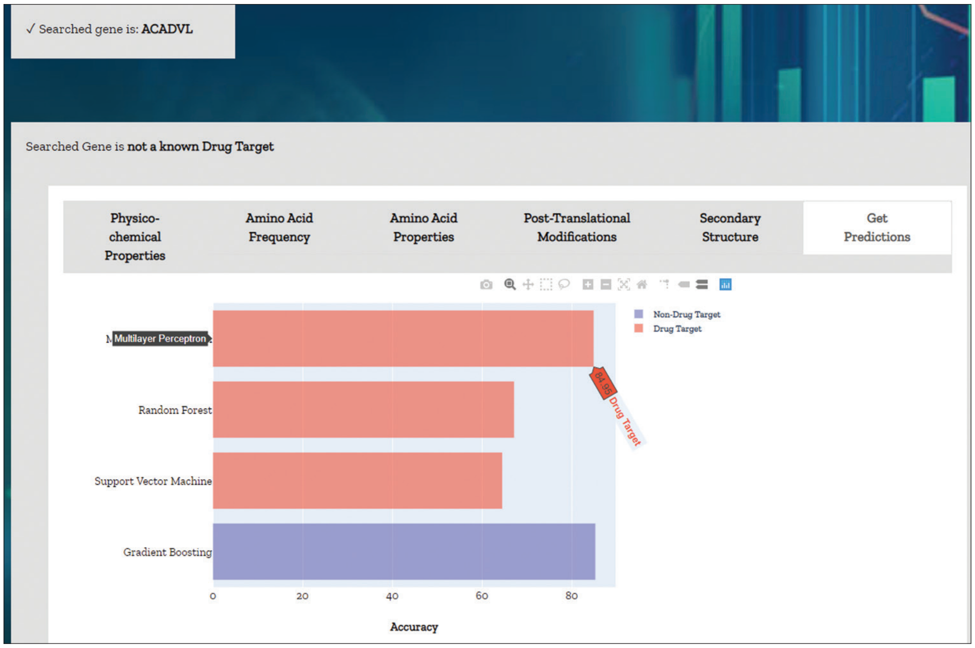
The search results page shows a data table which has a total of eight different tabs with information and plots. The first tab contains information regarding the druggable properties of the searched gene, along with information like the physico-chemical properties. The next four tabs contain different comparative plots for the values of features like amino acid frequency, frequency of different variants of amino acids (aromatic, aliphatic, basic, etc.,), post-translational modifications, and secondary structure. These plots are vertically stacked bar charts created using Plotly (Python's plotting library) and are used to compare the values of the searched protein with the average value of all drug target proteins and non-drug target proteins for a particular feature. The search results for the protein synaptic vesicle glycoprotein 2A (SV2A), a known AED target, are given in Figure 2.

Literature associated with epilepsy

The tab "Literatures Associated" is optional as they do not occur for every searched protein. It is displayed only if the searched protein has research literature in association with epilepsy, and the tab lists all those research articles. The epilepsy-related research articles are fetched from PubMed, a repository that hosts citations of 30 million biomedical articles.

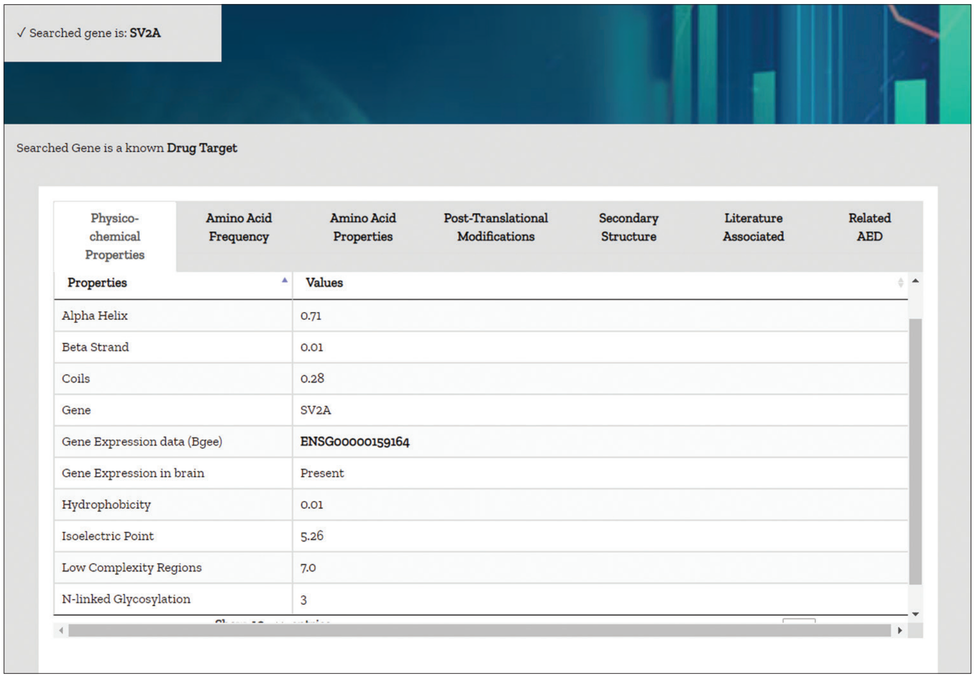
Related AEDs

The other tab called "Related AEDs" displays the anti-epileptic drugs whose target protein is the searched protein. If the searched protein is not a known drug target protein, this tab will not be visible since there will not be any existing AEDs for such a protein. The last tab is the "Get Predictions" tab which gives the machine learning predictions for the searched protein. There are predictions from four different classifiers stating whether the searched protein can be a probable drug target or not along with model accuracy. This tab is present only if the searched protein is not a known target protein.


✓ Searched gene is: **ACADVL**

Searched Gene is **not a known Drug Target**

Figure 1: Public domain databases used for building Epilepsy Database


✓ Searched gene is: **SV2A**

Searched Gene is a known **Drug Target**

Properties	Values
Alpha Helix	0.71
Beta Strand	0.01
Coils	0.28
Gene	SV2A
Gene Expression data (Bgee)	**ENSG00000159164**
Gene Expression in brain	Present
Hydrophobicity	0.01
Isoelectric Point	5.26
Low Complexity Regions	7.0
N-linked Glycosylation	3

Figure 2: Search results for the protein SV2A, a known AED target

Browsing through TREADS

There is a browse section in the navigation bar at the header of the website to browse AEDs, AED targets, miRNAs, and disorders [Figure 3]. The different sections of browse are as follows:

AEDs and AED targets

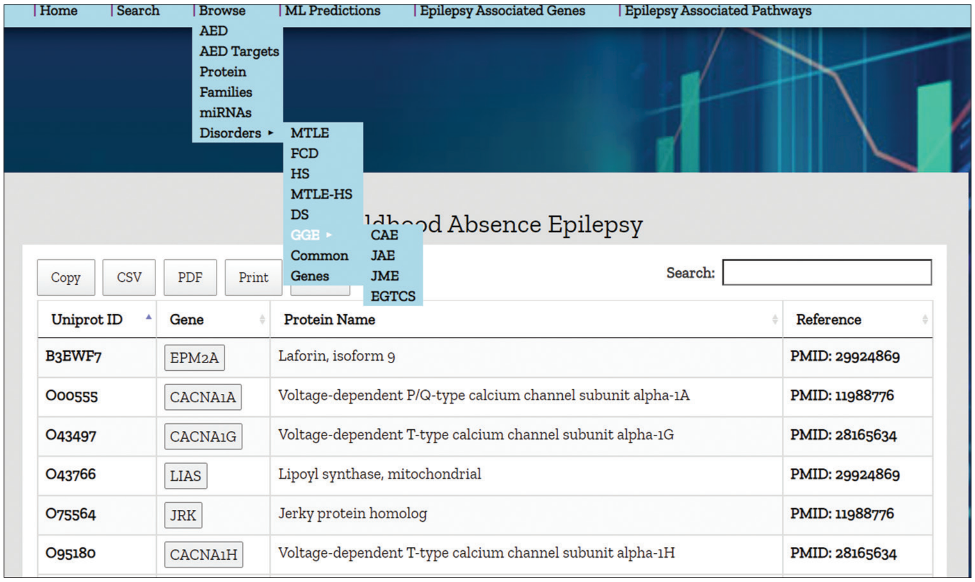
a. AED section displays all 46 drugs (approved, approved and investigational, investigational, and experimental and investigational) for epilepsy treatment [Figure 4]. The data is collected from the DrugBank repository and the corresponding links to the repository are provided. AED target section displays all 390 AED targets along with the corresponding drugs.

Protein families

b. This page shows the proportion of protein families that are associated with epilepsy. A pie chart is used to represent the information which is created using Plotly.

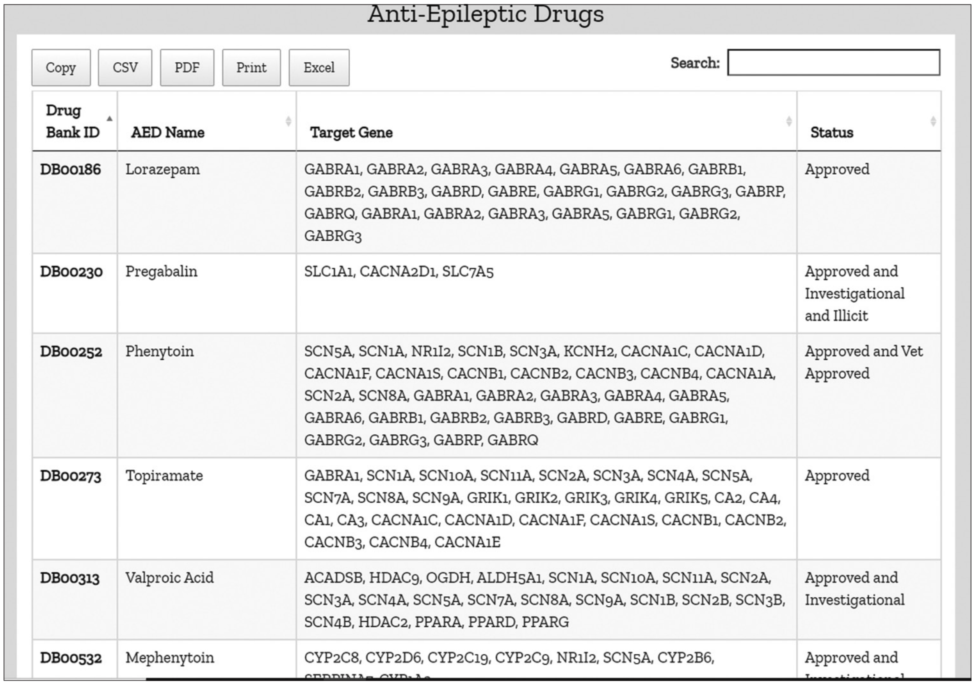
miRNAs

c. This page contains the experimentally studied 1987 miRNAs which interact with proteins that are found to be associated with epilepsy as per different literature articles. The type of experiment used,[13] and the corresponding research article is mentioned for every miRNA-protein pair.



Uniprot ID	Gene	Protein Name	Reference
B3EWF7	EPM2A	Laforin, isoform 9	PMID: 29924869
O00555	CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A	PMID: 11988776
O43497	CACNA1G	Voltage-dependent T-type calcium channel subunit alpha-1G	PMID: 28165634
O43766	LIAS	Lipoyl synthase, mitochondrial	PMID: 29924869
O75564	JRK	Jerky protein homolog	PMID: 11988776
O95180	CACNA1H	Voltage-dependent T-type calcium channel subunit alpha-1H	PMID: 28165634

Figure 3: Browse known AEDs, AED targets, epilepsy-associated protein families, miRNAs, and genes associated with epileptic disorders



Drug Bank ID	AED Name	Target Gene	Status
DB00186	Lorazepam	GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRA1, GABRA2, GABRA3, GABRA5, GABRG1, GABRG2, GABRG3	Approved
DB00230	Pregabalin	SLC1A1, CACNA2D1, SLC7A5	Approved and Investigational and Illicit
DB00252	Phenytoin	SCN5A, SCN1A, NR1I2, SCN1B, SCN3A, KCNH2, CACNA1C, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CACNA1A, SCN2A, SCN8A, GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ	Approved and Vet Approved
DB00273	Topiramate	GABRA1, SCN1A, SCN10A, SCN11A, SCN2A, SCN3A, SCN4A, SCN5A, SCN7A, SCN8A, SCN9A, GRIK1, GRIK2, GRIK3, GRIK4, GRIK5, CA2, CA4, CA1, CA3, CACNA1C, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CACNA1E	Approved
DB00313	Valproic Acid	ACADSB, HDAC9, OGDH, ALDH5A1, SCN1A, SCN10A, SCN11A, SCN2A, SCN3A, SCN4A, SCN5A, SCN7A, SCN8A, SCN9A, SCN1B, SCN2B, SCN3B, SCN4B, HDAC2, PPARA, PPARD, PPARG	Approved and Investigational
DB00532	Mephenytoin	CYP2C8, CYP2D6, CYP2C19, CYP2C9, NR1I2, SCN5A, CYP2B6,	Approved and

Figure 4: The table displaying known anti-epileptic drugs (AEDs) in TREADS

Disorders

d. This section contains different epileptic disorders, and within each disorder there are proteins which are associated to it. The corresponding reference to the research article stating the association is also provided. As of now, there are genes associated with nine different epileptic disorders, that is, mesial temporal lobe epilepsy (MTLE), focal cortical dysplasia (FCD), hippocampal sclerosis (HS), mesial temporal lobe epilepsy with hippocampal sclerosis (MTLE-HS), Dravet syndrome (DS), childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE), juvenile myoclonic epilepsy (JME), and generalized epilepsy with generalized tonic-clonic seizures (EGTCS). There is also a feature to find genes which are common across different disorders [Figure 5].

ML predictions

For machine learning–based predictions, a separate page is given in case the user wants to know whether a particular protein can be a probable drug target or not. This page lets the user search for a protein using the corresponding "Gene Name", and if the searched protein is a known drug target, the output is the search result page since for such proteins ML predictions are not required. In case, the searched protein is not a known drug target, the features of the protein are extracted and the possibility of the protein being druggable is displayed using ML-based models. ML predictions for the searched protein ACADVL (Very long-chain specific acyl-CoA dehydrogenase, mitochondrial, UniProtKB Identifier: P49748). Three of the classifiers predicted it to be a potential drug target and using one algorithm, it could not be predicted with high accuracy [Figure 6].

Epilepsy-associated genes and pathways

One thousand two hundred forty-four epilepsy-associated genes, and six different associated pathways reported in research articles are extracted and represented in tabular form. Each entry has corresponding references and links to the source of the data.

Discussion

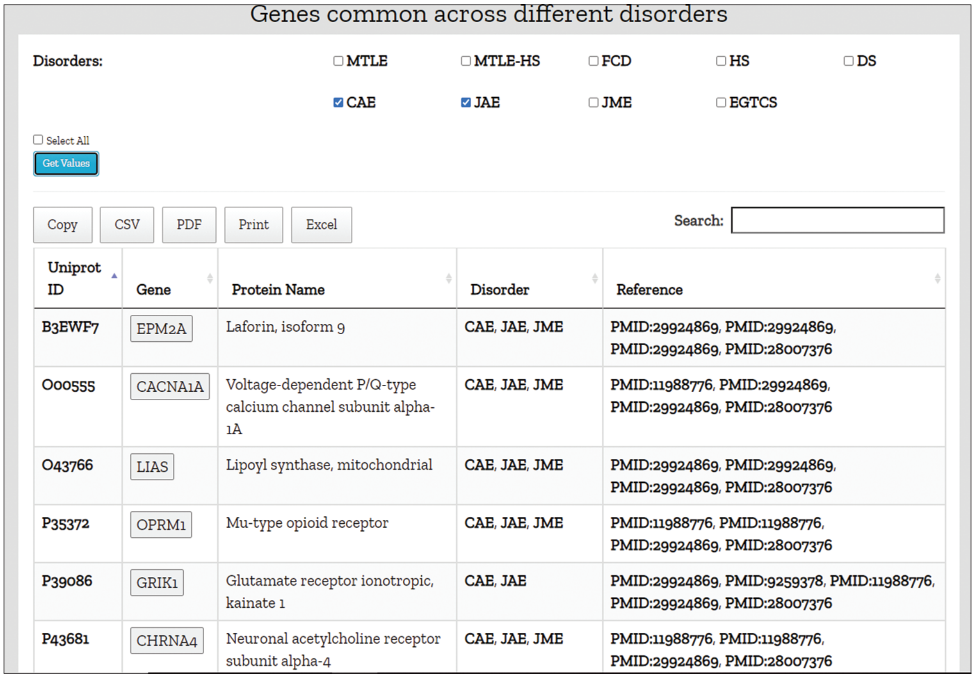
As epilepsy is one of the common neurological disorders with refractory seizures in nearly 30% patients, a platform for the research community to study anti-epileptic drugs and drug targets is needed. We developed a database with epilepsy-associated data integrated from various sources. The genes specific to a particular epileptic disorder and those common across multiple disorders can be browsed. TREADS also has ML classifier models to predict whether a protein can be a potential drug target or not. There are different features analyzed which contribute towards the druggability of a protein. Through the platform, the user can find if the protein of interest has druggable properties or not, and through comparative plots, can compare it with other proteins. TREADS provides a user-friendly UI to access all the data in tabular data, and the information provided in TREADS is downloadable in different formats, making it more useful. The platform can be accessed either from a desktop or a mobile device. Currently, the platform supports only the human reviewed genes from UniProt and cannot be searched using a protein sequence. Also, the information on variants within the genes or proteins, and the three-dimensional structural information is not provided. These will be included in the next version of TREADS.

Acknowledgements

Authors would like to thank the Executive director of C-DAC Bangalore for supporting this activity and for providing the required infrastructure to carry out the project.

Financial support and sponsorship

The project is financially supported by the Office of Principal Scientific Adviser (PSA), Govt. of India.



Genes common across different disorders

Disorders: □ MTLE □ MTLE-HS □ FCD □ HS □ DS ☑ CAE ☑ JAE □ JME □ EGTCS

□ Select All

Uniprot ID	Gene	Protein Name	Disorder	Reference
B3EWF7	EPM2A	Laforin, isoform 9	CAE, JAE, JME	PMID:29924869, PMID:29924869, PMID:29924869, PMID:28007376
O00555	CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A	CAE, JAE, JME	PMID:11988776, PMID:29924869, PMID:29924869, PMID:28007376
O43766	LIAS	Lipoyl synthase, mitochondrial	CAE, JAE, JME	PMID:29924869, PMID:29924869, PMID:29924869, PMID:28007376
P35372	OPRM1	Mu-type opioid receptor	CAE, JAE, JME	PMID:11988776, PMID:11988776, PMID:29924869, PMID:28007376
P39086	GRIK1	Glutamate receptor ionotropic, kainate 1	CAE, JAE	PMID:29924869, PMID:9259378, PMID:11988776, PMID:29924869, PMID:28007376
P43681	CHRNA4	Neuronal acetylcholine receptor subunit alpha-4	CAE, JAE, JME	PMID:11988776, PMID:11988776, PMID:29924869, PMID:28007376

Figure 5: Feature to search for genes that are associated with multiple epileptic disorders

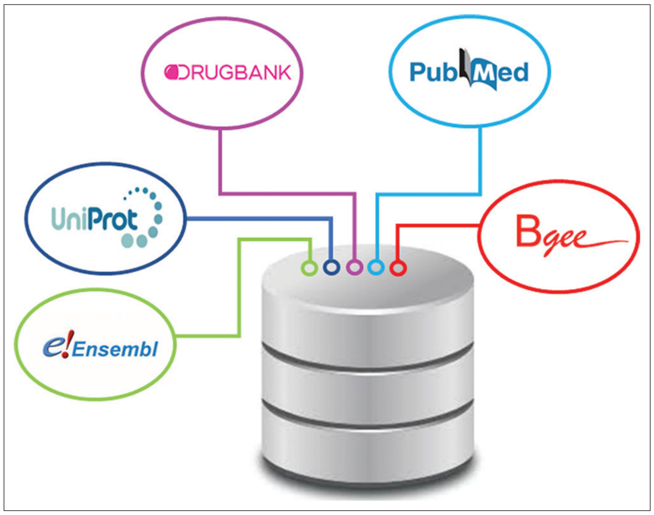


Figure 6: Results generated using machine learning classifier models for the searched protein ACADVL (UniProtKB Identifier: P49748)

Conflicts of interest

There are no conflicts of interest.

References

1. Dixit AB, Banerjee J, Chandra PS, Tripathi M. Recent advances in epilepsy research in India. Neurol India 2017;65(Suppl):S83-92.
2. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, *et al*. DrugBank 4.0: Shedding new light on drug metabolism. Nucleic Acids Res 2014;42:D1091-7.
3. Wang J, Lin Z-J, Liu L, Xu H-Q, Shi Y-W, Yi Y-H, *et al*. Epilepsy-associated genes. Seizure 2017;44:11-20.
4. Dixit AB, Banerjee J, Srivastava A, Tripathi M, Sarkar C, Kakkar A, *et al*. RNA-seq analysis of hippocampal tissues reveals novel candidate genes for drug refractory epilepsy in patients with MTLE-HS. Genomics 2016;107:178-88.
5. Cava C, Manna I, Gambardella A, Bertoli G, Castiglioni I. Potential role of miRNAs as theranostic biomarkers of epilepsy. Mol Ther Nucleic Acids 2018;13:275-90.
6. Ran X, Li J, Shao Q, Chen H, Lin Z, Sun ZS, *et al*. EpilepsyGene: A genetic resource for genes and mutations related to epilepsy. Nucleic Acids Res 2015;43:D893-9.
7. Klatt J, Feldwisch-Drentrup H, Ihle M, Navarro V, Neufang M, Teixeira C, *et al*. The EPILEPSIAE database: An extensive electroencephalography database of epilepsy patients. Epilepsia 2012;53:1669-76.
8. Basu SN, Kollu R, Banerjee-Basu S. AutDB: A gene reference resource for autism research. Nucleic Acids Res 2009;37:D832-6.
9. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics 2009;25:451-7.
10. Kumar A, Janaki C, Hosur MV, Pal SN. Machine learning techniques to identify potential drug targets for Anti-epileptic drugs. In: 2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT). 2020. p. 1-6.
11. UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49:D480-9.
12. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: Integrating and comparing heterogeneous transcriptome data among species. In: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Data Integration in the Life Sciences. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 124-31.
13. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, *et al*. miRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res 2020;48:D148-54.