PROMETEO 2024

INNOVATE IMPLEMENT IMPROVE

AI BioInnovate Challenge

PROBLEM STATEMENTS

AI-Bioinnovate Hackathon Problem Statements

- *Protein Engineering*

PE1 - Design and implement a model that can generate realistic and meaningful protein sequences based on a few given examples. Participants are provided with a small dataset of protein sequences for training, and they need to fine-tune a pre-trained language model or develop a novel model to achieve accurate and diverse protein sequence generation.

1. Dataset : dataset_ENZ

PE2 - Build a model for antitubercular peptides and find important physicochemical properties of aminoacids that have high correlation with activity of antitubercular peptides. Describes the physicochemical properties of amino acids using 566 indices (https://www.genome.jp/aaindex/AAindex/l); use these as features. (Encode amino acids with these properties.)

1. Apply feature selection methods to find non redundant features. Perform SHAP analysis to find key physicochemical properties.
2. Positive dataset: Positive_dataset
3. Negative dataset: Negative Dataset

PE3 - A model for predicting moon light proteins based on protein language models and deep learning.

1. Data: Files are in fasta format
2. Positive Data (positive data)
3. Negative Data (negative data)

PE4 - Identify features that significantly contribute to classification accuracy by applying various feature selection methods and eventually building a stacked ensemble method

1. The given dataset is a feature vector of blood brain barrier penetrating peptides. Class ‘1’ are blood brain barrier penetrating peptides; 0 are non-b3p2
2. It has different types of features. The final dataset has feature size greater than sample size. Provided training dataset (b3p2_training) and independent dataset (b3p2_fusion_independent) Training dataset

PE5 - Develop a model that can predict the immunogenicity of novel antigens for vaccine design.

1. Datasets- https://drive.google.com/drive/folders/1KOBj-cyRgybZQr5n6NfM_hYyUGBzX3n-?usp=drive_link

- *Multi-omic research*

MO1 - Develop an AI model that predicts potential drug-target interactions using multi-omic data (e.g., gene expression, protein-protein interactions, metabolic pathways) from various sources. Implement a model to predict drug-target interactions based on gene expression data.

MO2 - Construct a drug-disease network based on known drug-target interactions and disease-gene associations. Implement am algorithm to identify drugs with connections to multiple disease nodes.

MO3 - Develop machine learning models to analyze large chemical datasets and predict the potential effectiveness and safety of new drug candidates for specific medical conditions. Optimize drug discovery processes to reduce time and cost.

- *Small Molecules :*

SM1 - Build a Predictive model that predicts the pIC50 of the molecules. Select a library from the below links and filter them on the basis of pIC50 > 7 and follow all Lipinski's rule of 5.

Datasets:

a. Link 1
b. Link 2
c. Link 3

SM2 - Develop a predictive model that accurately identifies and predicts interaction between drugs and their target molecules, considering various biological, chemical, and structural features, to enhance drug discovery and development processes?

Link to datasets:

1. Link 1
2. Link 2
3. Link 3

SM3 - Develop a comprehensive toxicity prediction model that accurately assesses and predicts potential adverse effects of chemical compounds across diverse biological systems, including cell lines, animal models, and tissue models, incorporating relevant molecular, cellular, and organismal features to ensure robust toxicity evaluation in preclinical stages of drug development?

Datasets:

a. https://huggingface.co/datasets/zpn/clintox
b. https://paperswithcode.com/dataset/tox21-1
c. https://tdcommons.ai/single_pred_tasks/tox/

SM4 - Develop a methodology to systematically estimate and communicate uncertainties associated with predictions made by a deep learning model, taking into account sources such as limited data, model parameter uncertainty, and potential input variations, in order to enhance the trustworthiness and applicability of the model across different domains and scenarios?

Note: Consider 2 or 3 properties from the dataset provided below. (Both Classification and Regression)

Dataset:

https://tdcommons.ai/single_pred_tasks/adme/

SM5 - How can we devise a computationally efficient clustering approach that maintains high accuracy in grouping similar data points while minimizing the computational resources required, ensuring scalability for large datasets and real-time applications?

Dataset:

https://github.com/molecularsets/moses

- *Retrosynthesis*

RS1 - Create a model that could accurately predict the yield of a reaction; it should be able to accurately predict the yield outcome for all diverse reaction classes.

RS2 - Build a seq to seq based model which could predict the reaction conditions for the reaction i.e. given reaction smiles as an input it should be able to predict the conditions.

choriso dataset for problems 17 and 18

RS3 - Create a multistep algorithm which could identify the potential retrosynthetic pathways that terminate at building blocks for a single step prediction you could use the pretrained model like Local Retro. It should be able to predict more pathways for a given molecule.

building blocks

Clinical Trials

CT1 - Clinical trial outcome prediction: prediction of clinical trial success. Use explainable AI to explain how you predict the outcome

i. Data set : https://clinicaltrials.gov
ii. https://www.cancer.gov/ccg/research/genome-sequencing/tcga

CT2 - Disease Progression prediction:Predict the progression of biomarkers of the disease over the years.

iii. Data: For alzhemier disease:https://adni.loni.usc.edu/
iv. https://naccdata.org/
v. Beat Acute Myeloid Leukemia (AML) 1.0
vi. https://www.ppmi-info.org/access-data-specimens/download-data-parkinson's disease
vii. https://www.oasis-brains.org/

CT3 - Develop a model that utilizes machine learning to predict Adverse events

Dataset : https://clinicaltrials.gov/