# Making Sense of Missense: Machine Learning Training On a Model for Missense Variants Data

Margaux Vasilescu, Bronx High School of Science, New York City

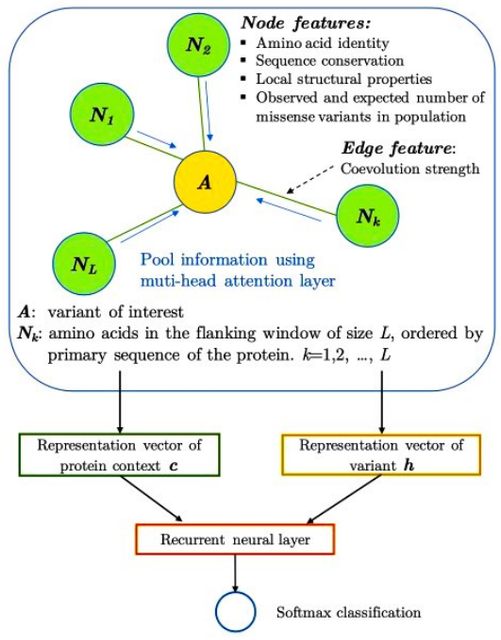
Project ID: CELL-415

## Question 1: Research Question

- Can a **machine learning model** that analyzes **missense variants** data be improved by training it with different data?
- In particular, can data from another machine learning program, **MSA Transformer**, be fed into and train gMVP model to make it more efficient?

**Missense variants** are genetic mutations that can alter the function of the protein.

Some **missense variants** cause terrible diseases: **cystic fibrosis, sickle cell anemia, certain cancers**

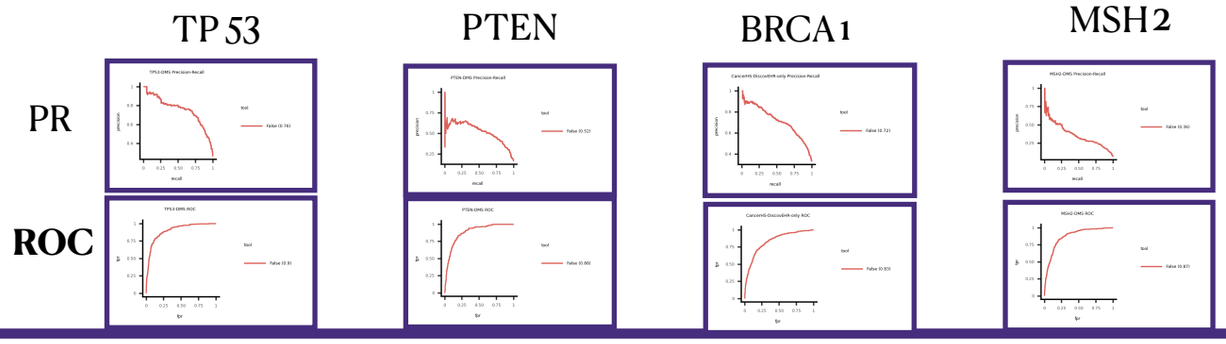



**gMVP machine learning**

## Question 3: Data Analysis & Results

| Training Data | TP53 ROC | TP53 PR | PTEN ROC | PTEN PR | BRCA1 ROC | BRCA1 PR | MSH2 ROC | MSH2 PR |
|---|---|---|---|---|---|---|---|---|
| Original | 0.87 | 0.71 | 0.87 | 0.55 | 0.83 | 0.71 | 0.86 | 0.33 |
| MSA Trans-former | 0.9 | 0.76 | 0.88 | 0.52 | 0.83 | 0.72 | 0.87 | 0.36 |

**MSA Transformer** Data improves gMSV training scores compared to original data




## Question 2: Methodology

1. Install PyTorch machine learning program and related Python libraries.
2. Download **gMVP** from Github.
2. Train gMVP with original data on workstation with GPU.
3. At end of training, gMVP creates **AUC-ROC** and **Precision Recall** plots for 4 types of disease related missense variants: **TP53, PTEN, BRCA1, MSH2**

GPU

1. Edit **gMVP** program code to access **MSA Transformer Data.**
2. Edit code so the Original Data file names match up with and access the related MSA Transformer Data,
3. If there is no related MSA Transformer Data for the original data file, create a tensor with zeros shaped with vectors [129, 235]
4. If MSA Transformer data exists for the original data file, reshape two types of MSA Transformer tensor files: "row_attention.pt and contact.pt
5. row_attention.pt tensor reshaped from original shape of 5 vector matrix of [1, 12, 12, seq, seq] to 2 vector shape of [129, 235]
6. contact.pt tensor reshaped from original shape of 3 vector matrix of [1, seq, seq] to 2 vector sequence of [129. 235]
7. Edit program to concatenate row_attention tensor with contact tensor
8. Have program create a window of 128 AAs around the variant position
9. Feed the new concatenated row_attention/contact tensor as the "pairwise" in the program

**At end of MSA Transformer data training**, **gMVP** creates **AUC-ROC** and **Precision Recall** plots for 4 types of disease related missense variants: **TP53, PTEN, BRCA1, MSH2**

## Question 4: Interpretations & Conclusions

- **Result table shows that the MSA Transformer data improved the ROC for 3 of the 4 types of diseases: TP53, PTEN and MSH2.**
- **For BRCA1, the ROC for MSA Transformer data training had the same ROC as the Original Data.**
- **For Precision-Recall, the MSA Transformer data improved 3 of 4 types of diseases: TP53, BRCA1 and MSH2.**
- **Compared to Original Data, for MSA Transformer, Precision-Recall score reduced for PTEN.**
- **Overall, MSA Transformer Data improved the training of the gMVP model to make it more efficient in finding missense variants that changes protein and creates diseases.**
- **One flaw in my experiment:** my run or Original Data training gMVP Model had scores lower than the scores in the original article for gMVP by those who created gMVP. Need to run the program again and find why the Original Data scores are lower than the gMVP article.
- **Overall, initial research shows that MSA Transformer Data improves the gMVP Model**