# **CAFA 5 Protein Function Prediction**

Solution Introduction





1 Data Engineering

2 CV Strategy



### **Architecture Preview**



Linear with T5
Sequence Embedding

Linear with ESM2
Sequence Embedding

Linear with ProtBERT Sequence Embedding

CLIP-based model with MPNet
Sentence Transformer

CLIP-based model with BioBERT



Ensemble

1 Data Engineering

2 CV Strategy



# Data Engineering – Linear Regression Based Models



#### Sequence-side Features

- 1. T5 / ESM2 / ProtBERT feature vector
- 2. length of sequence
- 3. protein structure feature
- 4. mean & std. of amino acid property feature
- 5. ratio of each amino acids in sequence
- 6. ratio of each amino acids' group in sequence
- 7. taxonomic identifier

# Data Engineering – CLIP Based Models



#### Sequence-side Features

- 1. T5 feature vector
- 2. length of sequence
- 3. protein structure feature
- 4. mean & std. of amino acid property feature
- 5. ratio of each amino acids in sequence
- 6. ratio of each amino acids' group in sequence
- 7. taxonomic identifier

#### **GO Term-side Features**

- 1. MPNet(ST) / BioBERT feature vector
- 2. GO Term type
- 3. word2vec embedding
- 4. GNN embedding

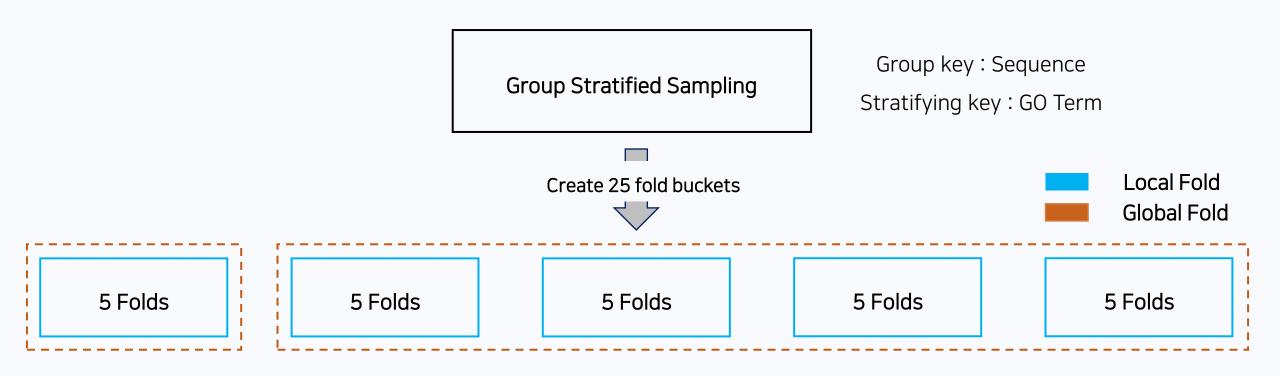
1 Data Engineering

2 CV Strategy



# CV Strategy – Splitting Dataset





Linear-based model uses Local Fold & CLIP-based model use Global Fold



Twin CV allow to train linear-based model with data representing same distribution as total data

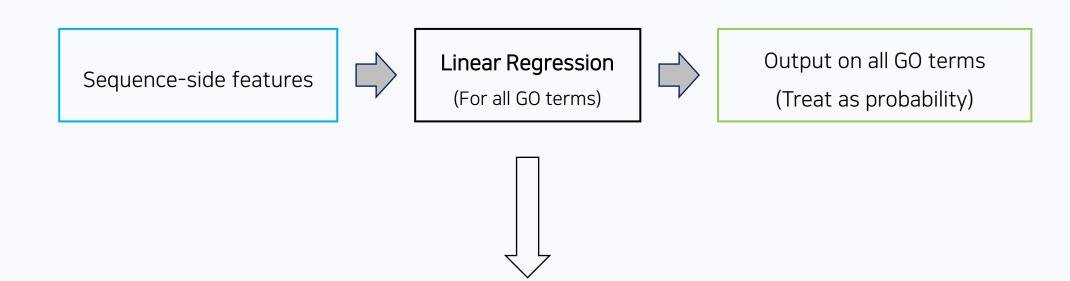
1 Data Engineering

2 CV Strategy



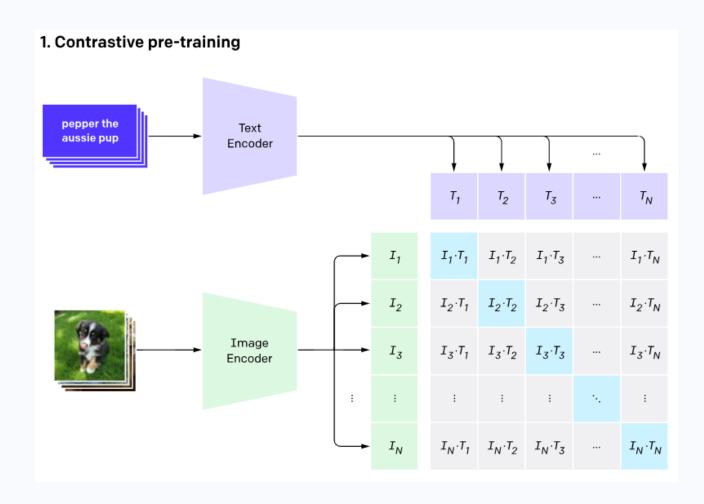
# Architecture – Linear Regression Based Models





This simple model allow high speed training for all GO Terms



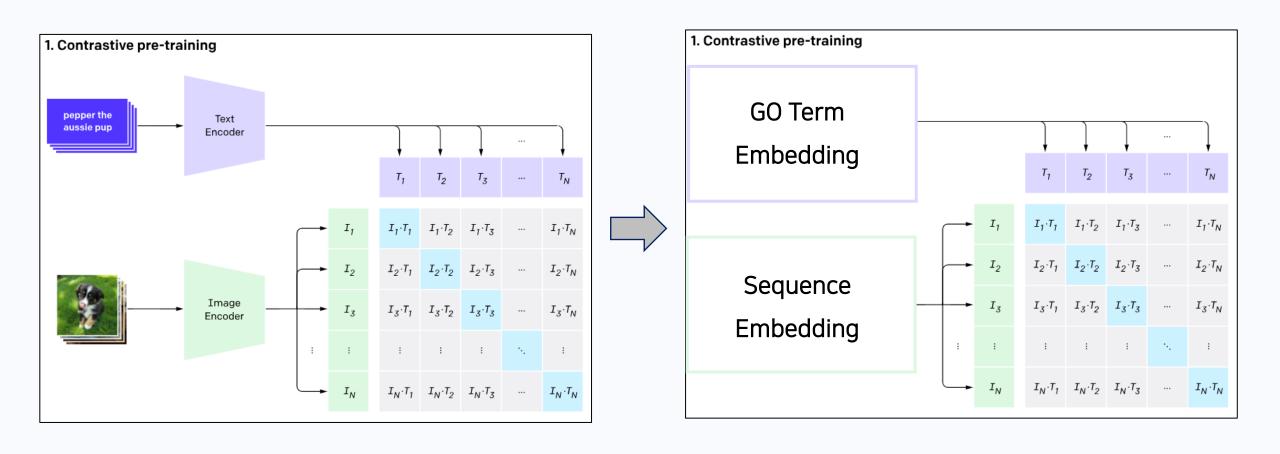


- 1 Using both query & content sides embedding vector
- 2 Calculating output with matrix multiplication



Matrix multiplication allow to calculate probability on all contents from a input query with fast speed

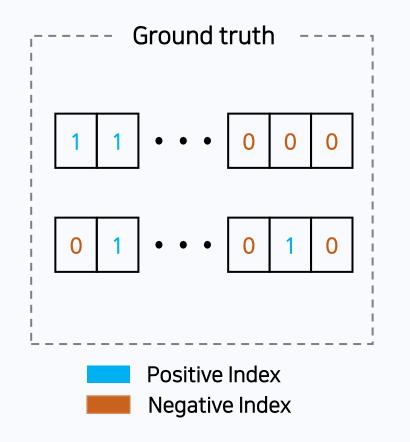




But, what do we do to migrate original model when it is not a multiclass classification task?



#### Introduction on **Dynamic Negative Sampling** technique for binary classification

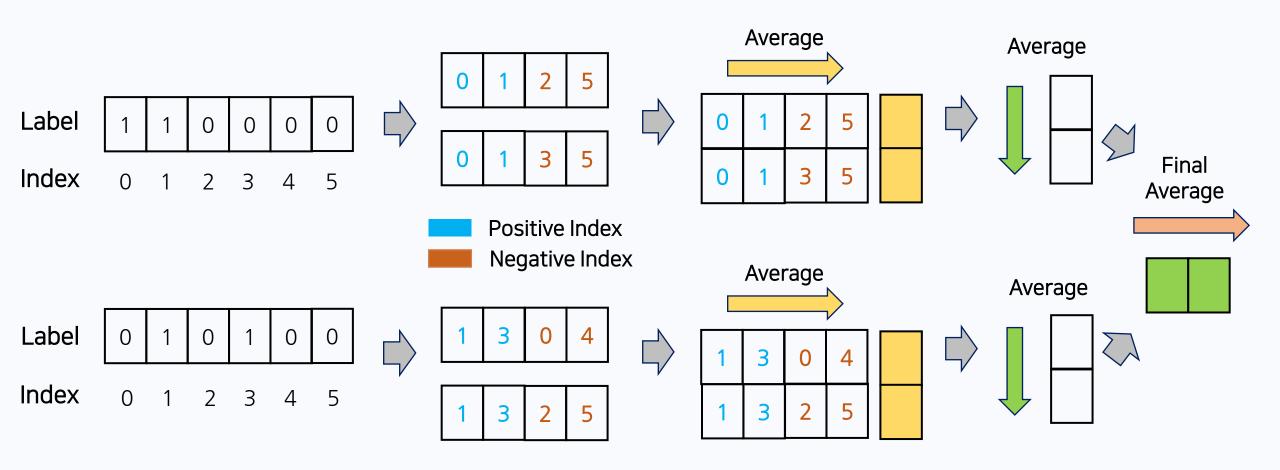




# Calculating loss with Dynamic Negative Sampling Process

- 1. Shuffling negatives
- 2. Select number of negatives (\*negative\_sampling\_ratio)
- 3. Select number of combinations (\*n\_combinations)
- 4. Average on elements' loss in each combinations
- 5. Average on all combinations' loss
- 6. Average on all batches' loss





# Architecture - Ensemble



0.20

Linear with T5
Sequence Embedding

0.15

Linear with ESM2
Sequence Embedding

0.15

Linear with ProtBERT Sequence Embedding

Weight

0.25

CLIP-based model with MPNet
Sentence Transformer

0.25

CLIP-based model with BioBERT



Weighted Average Ensemble

# Architecture - Evaluation Score



0.49170

Linear with T5
Sequence Embedding

0.47478

Linear with ESM2
Sequence Embedding

0.45420

Linear with ProtBERT Sequence Embedding

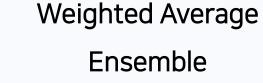
PB Score

0.50111

CLIP-based model with MPNet
Sentence Transformer

0.49750

CLIP-based model with BioBERT



0.53672



# Thank You

The End



# Appendix – Public Model Evaluation Score



