PETA: evaluating the impact of protein transfer learning with sub-word tokenization on downstream applications

Journal of Cheminformatics, 2024

LifeLU Reading Club

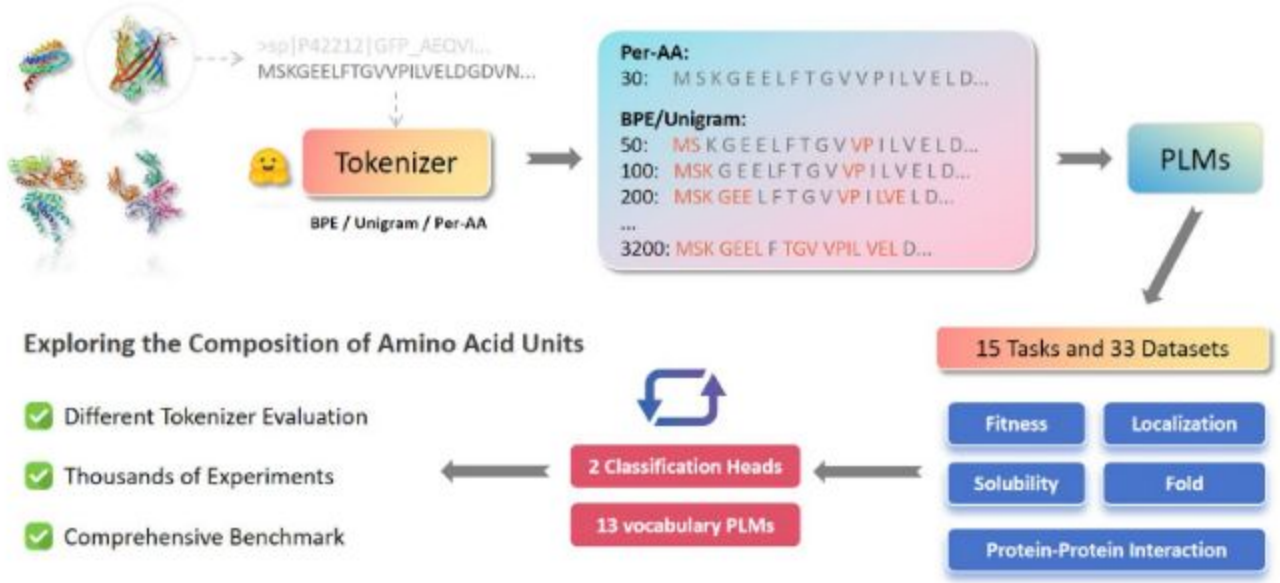
Amirreza Sattarzadeh - 17 July 2025
(Emir Rıza Settarzade)

Introduction: Proteins - The Building Blocks of Life

- **Crucial Role in Biology & Medicine:**
  - Genetic Engineering
  - Drug Discovery
  - Enzyme Catalysis
- **Challenges in Protein Engineering:**
  - **Traditional Lab-based methods:**
    - Time-consuming & Costly
    - High expertise dependency
  - **Computational methods:**
    - Often lack sufficient accuracy

The Rise of Protein Language Models (PLMs)

- **Powerful Tools:** PLMs excel in learning protein representations.
- **Traditional Tokenization:**
  - Proteins are sequences of ~20 amino acids.
  - PLMs traditionally tokenize each amino acid as a separate token.
  - **Vocabulary Size:** ~20 (very small compared to NLP models)
- **The Challenge:**
  - This ignores **frequently occurring amino acid combinations** (sub-sequences / "protein words").
  - Viewing individual amino acids as isolated tokens is inefficient and insufficient for capturing complex biological information.


>sp|P42212|GFP_AEQVI...
MSKGEELFTGVVPILVELDGDVN...
Tokenizer
BPE / Unigram / Per-AA
Per-AA:
30: M S K G E E L F T G V V P I L V E L D...
BPE/Unigram:
50: MS K G E E L F T G V VP I L V E L D...
100: MSK G E E L F T G V VP I L V E L D...
200: MSK GEE L F T G V VP I LVE L D...
...
3200: MSK GEEL F TGV VPIL VEL D...
PLMs
15 Tasks and 33 Datasets
Fitness
Localization
Solubility
Fold
Protein-Protein Interaction
2 Classification Heads
13 vocabulary PLMs
Exploring the Composition of Amino Acid Units
Different Tokenizer Evaluation
Thousands of Experiments
Comprehensive Benchmark

The Solution: Sub-word Tokenization

Inspired by Natural Language Processing (NLP):

- Algorithms like BPE (Byte-Pair Encoding) and Unigram have revolutionized NLP.

Our Approach:

- **BPE:** Iteratively merges the most frequent adjacent character/token pairs.
  - *In proteins: helps identify common motifs/domains.*
- **Unigram:** Builds a vocabulary by selecting the most probable token combinations based on frequency.
  - *In proteins: can identify rare but important amino acid patterns.*

Goal: Develop a universal amino acid encoding approach for robust performance across diverse protein-related tasks.

PETA: The Comprehensive Benchmark

- **What is PETA?**

  A comprehensive and systematic benchmark for evaluating PLMs.

- **Scope:**

  33 datasets categorized into **15 distinct downstream tasks**.

- **5 Main Task Categories:**

  Protein Fitness Prediction: Assessing fitness scores across mutations.

  Protein-Protein Interaction (PPI) Prediction: Identifying interactions between protein pairs.

  Protein Localization Prediction: Predicting cellular location of proteins.

  Protein Solubility Prediction: Predicting how well a protein dissolves.

  Protein Fold Prediction: Predicting the 3D structural class of a protein.

- **Focus:** Primarily on **protein-wise** and **protein-pair** tasks.

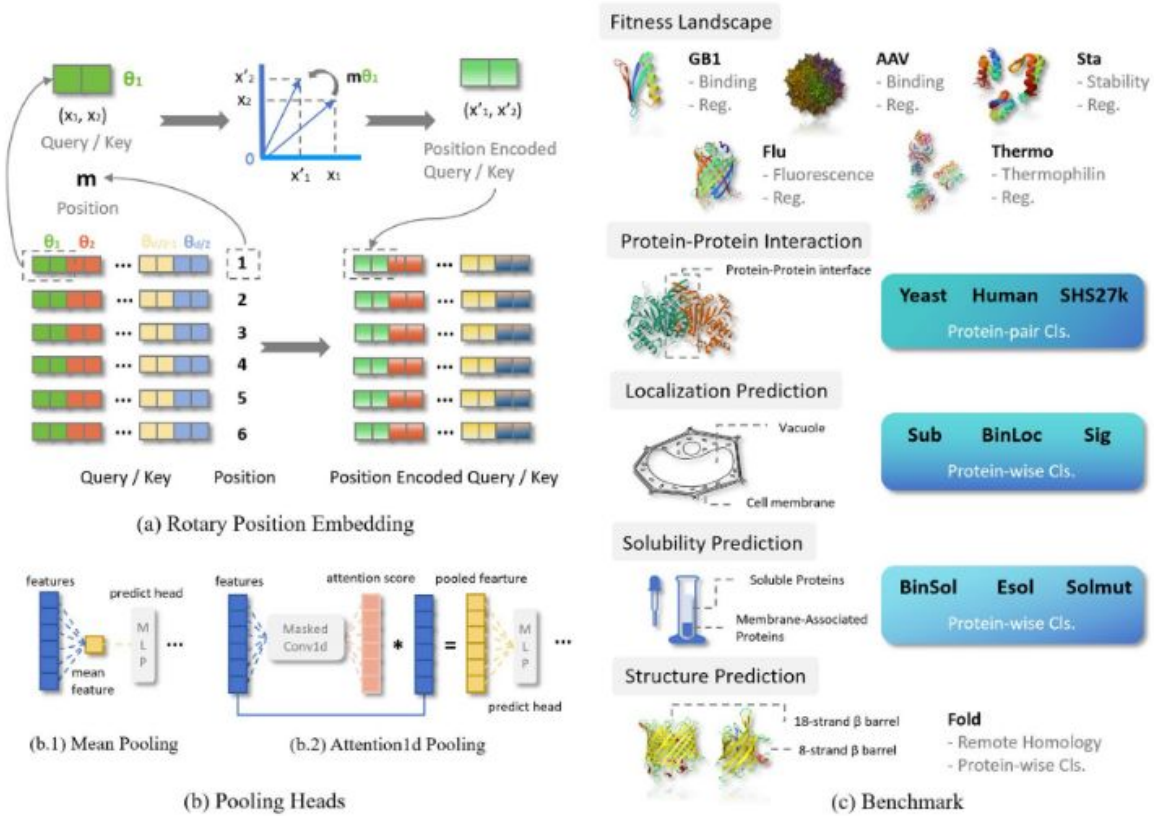


Fig. 2 The framework of PETA. (a) Pre-trained models use rotary position embedding, which possesses favorable theoretical properties and is an absolute positional encoding applicable to linear Attention. (b) We employed two distinct classification heads, namely mean pooling and attention1d pooling. The former is the most commonly used method at present, while the latter is relatively more complex. (c) Our benchmark comprises 15 downstream tasks, which can be categorized into five groups. Some of these downstream tasks include multiple datasets or data splitting methods, amounting to a total of 33 datasets

Table 1 Benchmark task details. Each task, along with its task name, category, the count of datasets or splits, the source of the dataset, and evaluation metric are shown below

Task name	Category	Count	Source	Metric
Fitness Prediction				
GB1 fitness (GB1)	Reg.	5	FLIP [70]	Spearman's ρ
AAV fitness (AAV)	Reg.	7	FLIP [70]	Spearman's ρ
Thermostability (Thermo)	Reg.	3	FLIP [70]	Spearman's ρ
Fluorescence (Flu)	Reg.	1	TAPE [71]	Spearman's ρ
Stability (Sta)	Reg.	1	TAPE [71]	Spearman's ρ
Protein-Protein Interaction Prediction				
Yeast PPI (Yeast)	Cls.	1	PETA	Accuracy
Human PPI (Human)	Cls.	1	PETA	Accuracy
SHS PPI (SHS27k)	Cls.	1	PETA	Accuracy
Localization Prediction				
Subcellular localization (Sub)	Cls.	3	pro-loc [83], DeepLoc-2 [75]	Accuracy
Binary localization (BinLoc)	Cls.	1	pro-loc [83]	Accuracy
Sorting signal (Sig)	Cls.	1	DeepLoc-2 [75]	Accuracy
Solubility Prediction				
Binary solubility (BinSol)	Cls.	1	DeepSol [84]	Accuracy
E.coli solubility (Esol)	Reg.	1	GraphSol [85]	MSE
Mutation solubility (Solmut)	Reg.	3	PETA	Spearman's ρ
Fold Prediction				
Fold Prediction (Fold)	Cls.	3	TAPE [71]	Accuracy

Reg.: regression; Cls.: classification; MSE: mean square error; Spearman's ρ: Spearman Correlation

Methods: Model Pipeline

Base Architecture: RoFormer

- An **autoencoding model** (similar to BERT)
- Enhanced with **Rotary Positional Embeddings (RoPE)**
    - *Improves understanding of sequential relationships & long-range dependencies.*

Pre-training Objective: Masked Language Modeling (MLM)

- Randomly mask tokens, then predict them based on context.
- *Helps learn the "language of proteins" and rich representations.*

Pre-training Data: UniRef90

- Comprehensive protein sequence database (~138 million sequences).
- *Crucial for learning general protein patterns.*

Methods: Amino Acid Segmentation & Vocabulary Size

- **Three Tokenization Methods Examined:**
  - **Per-Amino-Acid (Per-AA):**
    - Baseline method: Each amino acid is a separate token. (20 tokens)
    - *Allows high-resolution analysis of subtle changes.*
  - **Byte-Pair Encoding (BPE):**
    - Merges most frequent adjacent token pairs.
    - *Identifies common protein motifs and structural domains.*
  - **Unigram Language Modeling:**
    - Selects most probable token combinations based on frequency.
    - *Identifies rare but important amino acid patterns.*
- **Vocabulary Sizes Investigated:**
  - 50, 100, 200, 800, 1600, 3200 elements.
  - *A key variable to understand optimal representation.*

Methods: Classification Heads & Evaluation Metrics

Classification Heads (for Downstream Tasks):

- **Mean Pooling:**
  - Standard method, averages features for prediction.
- **Attention1d Pooling:**
  - More complex, uses attention mechanism to assign different weights to parts of the sequence.
  - *Often yields better performance.*

Evaluation Metrics (Task-dependent):

- **Spearman's ρ (Rho):**
  - For **Regression Tasks** (e.g., Fitness, Stability).
  - Measures **rank correlation** between predicted and true values.
- **Accuracy:**
  - For **Classification Tasks** (e.g., PPI, Localization, Fold Prediction).
  - Ratio of correct predictions to total predictions.
- **MSE (Mean Squared Error):**
  - For specific regression tasks (less common in this study's primary metrics).

Crucial for unbiased evaluation:

- Used both tokenization methods (BPE & Unigram), two pooling heads, and three random seeds for each vocabulary size.

Results: Significant Impact of Vocabulary Size

- **Profound Influence:** Vocabulary size profoundly impacts protein representation quality.
- **"Bell-shaped" Curve:** Performance generally shows a bell-shaped curve:
  - **Increases** with vocabulary size initially.
  - **Decreases** after an optimal point.
- **Optimal Threshold:** Models with **50 and 200 elements** achieved **optimal performance**.
- **Detrimental Effects:** Vocabulary sizes exceeding **800 elements** often led to **worse performance** compared to the Per-AA baseline (20 tokens).
  - *Unlike NLP, where larger vocabularies are usually better.*

Table 4 The number of datasets or splits whose average score exceeds the baseline model of 20 vocabulary size

Vocabulary	Sum (33)	Fit (17)	PPI (3)	Loc (5)	Sol (5)	Fold (3)
50	22	10	3	5	4	0
100	19	8	2	5	4	0
200	20	9	3	4	4	0
800	16	5	3	4	4	0
1,600	15	5	2	4	4	0
3,200	15	5	2	4	4	0

Fit: protein fitness; PPI: protein-protein interaction; Loc: protein localization; Sol: protein solubility; Fold: protein fold

Table 5 Performance on Fitness Prediction and Localization Prediction

Vocabulary	Fitness Prediction						Localization Prediction			
	GB1(5) ↑	AAV(7) ↑	Thermo(3) ↑	Flu(1) ↑	Sta(1) ↑	Sub-1(1) ↑	BinLoc(1) ↑	Sub-2(1) ↑	Sub-hpa(1) ↑	Sig(1) ↑
20	$48.6_{(0.8)}$	$34.8_{(0.3)}$	$61.3_{(0.2)}$	$39.3_{(1.0)}$	$\mathbf{51.3}_{(0.3)}$	$\mathbf{94.6}_{(0.1)}$	$91.3_{(0.3)}$	$92.2_{(0.1)}$	$89.0_{(0.1)}$	$95.7_{(0.1)}$
50	$\mathbf{50.3}_{(0.9)}$	$31.8_{(0.8)}$	$61.8_{(0.2)}$	$39.9_{(2.3)}$	$53.0_{(0.5)}$	$94.8_{(0.0)}$	$91.7_{(0.3)}$	$92.6_{(0.0)}$	$\mathbf{89.1}_{(0.2)}$	$96.1_{(0.0)}$
100	$53.4_{(0.4)}$	$\mathbf{31.3}_{(0.2)}$	$62.1_{(0.1)}$	$38.3_{(2.0)}$	$51.6_{(0.7)}$	$94.7_{(0.1)}$	$91.3_{(0.1)}$	$92.3_{(0.0)}$	$89.0_{(0.0)}$	$\mathbf{95.9}_{(0.2)}$
200	$50.4_{(0.1)}$	$30.6_{(1.0)}$	$\mathbf{61.6}_{(0.1)}$	$\mathbf{42.0}_{(1.9)}$	$49.5_{(0.5)}$	$94.5_{(0.1)}$	$\mathbf{91.3}_{(0.2)}$	$\mathbf{92.2}_{(0.0)}$	$89.4_{(0.0)}$	$95.9_{(0.1)}$
800	$46.2_{(1.2)}$	$28.8_{(0.2)}$	$60.6_{(0.3)}$	$41.5_{(1.7)}$	$46.4_{(0.1)}$	$94.4_{(0.1)}$	$90.9_{(0.3)}$	$92.1_{(0.0)}$	$89.4_{(0.1)}$	$95.6_{(0.1)}$
1,600	$46.5_{(0.9)}$	$27.4_{(0.1)}$	$61.3_{(1.0)}$	$43.8_{(0.9)}$	$43.2_{(1.6)}$	$94.4_{(0.1)}$	$90.5_{(0.2)}$	$92.1_{(0.0)}$	$89.2_{(0.1)}$	$95.6_{(0.1)}$
3,200	$46.9_{(1.1)}$	$29.9_{(0.9)}$	$60.8_{(0.1)}$	$\mathbf{43.0}_{(1.9)}$	$44.5_{(1.0)}$	$94.3_{(0.2)}$	$90.9_{(0.1)}$	$92.1_{(0.0)}$	$89.0_{(0.1)}$	$95.6_{(0.2)}$

Each value indicates the $mean_{(std)}$ score across all experiments within the same vocabulary size. The values colored with are higher than the Per-AA method. Datasets marked with (*) indicate the number of dataset splits. For instance, **GB1** encompasses five different dataset splits within the same dataset. The score with a vocabulary size of 50 reflects results across 60 experiments (5×2×2×3, representing the number of dataset splits, tokenization methods, classification heads, number of random seed experiments)

† The top three are highlighted by First, Second, Third.

Results: Performance Across Specific Tasks

Protein Fold Prediction:

- **Inverse relationship:** Increasing vocabulary size led to *decreased* performance.
- *Suggests larger vocabularies might obscure specific positional information.*

Protein Localization Prediction:

- Consistently high classification accuracy (>90%).
- **Minimal impact:** Performance variations remained within ~1%, regardless of vocabulary size.
- *Implies these features are robustly captured irrespective of token granularity.*

Protein Fitness Prediction:

- Improved with vocabulary expansion in most cases.
- **Exceptions:** AAV (significant drop), GB1 & Stab (declined after 200 tokens).

Protein-Protein Interaction (PPI) Prediction:

- Clear **performance improvement** with larger vocabularies.
- Higher gains in "harder to classify" datasets (e.g., Yeast, SHS27k).

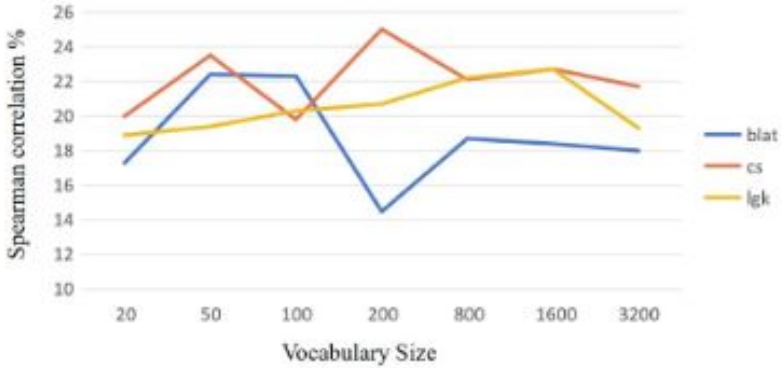


Fig. 4 A detailed exposition is provided on the performance results of three distinct protein solubility mutation datasets: Beta-lactamase TEM (blat), Chalcone Synthase (cs), and Levoglucosan Kinase (lgk) across varying vocabulary sizes

Table 6 Performance on PPI Prediction, Solubility Prediction and Fold Prediction

Vocabulary	PPI Prediction			Solubility Prediction			Fold Prediction		
	Yeast(1) ↑	SHS27k(1) ↑	Human(1) ↑	Esol(1) ↓	BinSol(1) ↑	Solmut(3) ↑	superfamily(1) ↑	family(1) ↑	fold(1) ↑
20	$69.9_{(0.7)}$	$51.9_{(0.8)}$	$92.7_{(0.2)}$	$0.044_{(0.008)}$	$67.0_{(0.5)}$	$18.7_{(0.7)}$	$51.4_{(0.4)}$	$93.0_{(0.5)}$	$26.7_{(0.4)}$
50	$72.5_{(0.7)}$	$52.6_{(0.2)}$	$93.0_{(0.2)}$	$0.045_{(0.002)}$	$69.1_{(0.1)}$	$21.7_{(0.7)}$	$50.5_{(0.1)}$	$92.9_{(0.2)}$	$26.8_{(0.6)}$
100	$72.0_{(0.9)}$	$52.0_{(0.3)}$	$92.5_{(0.2)}$	$0.048_{(0.000)}$	$69.6_{(0.3)}$	$20.8_{(0.3)}$	$43.7_{(0.7)}$	$91.9_{(0.3)}$	$22.6_{(0.6)}$
200	$72.8_{(0.5)}$	$53.5_{(0.3)}$	$92.9_{(0.1)}$	$0.047_{(0.001)}$	$70.2_{(0.1)}$	$20.1_{(1.4)}$	$43.9_{(0.4)}$	$91.5_{(0.2)}$	$26.4_{(1.9)}$
800	$72.0_{(0.6)}$	$53.4_{(0.3)}$	$92.8_{(0.2)}$	$0.046_{(0.001)}$	$70.2_{(0.7)}$	$21.0_{(0.7)}$	$43.5_{(0.5)}$	$89.4_{(0.5)}$	$24.0_{(0.2)}$
1,600	$73.3_{(0.6)}$	$53.1_{(0.7)}$	$92.5_{(0.6)}$	$0.048_{(0.001)}$	$70.4_{(1.3)}$	$21.3_{(1.3)}$	$41.8_{(0.5)}$	$89.0_{(0.4)}$	$23.1_{(1.4)}$
3,200	$72.4_{(0.5)}$	$52.0_{(0.2)}$	$92.2_{(0.2)}$	$0.047_{(0.000)}$	$69.9_{(0.9)}$	$19.7_{(1.1)}$	$40.6_{(0.2)}$	$88.0_{(0.4)}$	$23.3_{(0.4)}$

Each value indicates the $mean_{(std)}$ score across all experiments within the same vocabulary size. The values colored with are higher than the Per-AA method. Datasets marked with (*) indicate the number of dataset splits

† The top three are highlighted by First, Second, Third.

Table 7 The average results of different downstream task groups under the same vocabulary with varying tokenization methods are presented

Vocab.	BPE					Unigram				
	Fit	PPI	Loc	Sol	Fold	Fit	PPI	Loc	Sol	Fold
20	**47.1**$_{(0.2)}$	71.5$_{(0.5)}$	**93.2**$_{(0.1)}$	42.9$_{(0.2)}$	**57.2**$_{(0.1)}$	**47.1**$_{(0.1)}$	71.5$_{(0.5)}$	**93.2**$_{(0.1)}$	42.9$_{(0.2)}$	**57.2**$_{(0.2)}$
50	**47.1**$_{(0.1)}$	72.2$_{(0.2)}$	**93.4**$_{(0.2)}$	**45.6**$_{(0.3)}$	**55.8**$_{(0.1)}$	**47.7**$_{(0.3)}$	**73.1**$_{(0.5)}$	**93.3**$_{(0.1)}$	45.3$_{(1.0)}$	**57.6**$_{(0.3)}$
100	**47.1**$_{(0.2)}$	72.1$_{(0.3)}$	**93.2**$_{(0.1)}$	44.6$_{(0.2)}$	53.4$_{(0.6)}$	**47.6**$_{(0.2)}$	72.2$_{(0.4)}$	93.0$_{(0.1)}$	**45.8**$_{(0.8)}$	52.1$_{(0.2)}$
200	**47.4**$_{(0.5)}$	**73.1**$_{(0.1)}$	**93.2**$_{(0.2)}$	**45.4**$_{(1.1)}$	**54.9**$_{(0.8)}$	46.2$_{(0.3)}$	**73.1**$_{(0.5)}$	**93.1**$_{(0.1)}$	44.9$_{(0.4)}$	**53.0**$_{(0.4)}$
800	45.4$_{(0.5)}$	**72.8**$_{(0.5)}$	93.0$_{(0.2)}$	**45.6**$_{(0.3)}$	52.3$_{(0.2)}$	44.1$_{(0.1)}$	**72.6**$_{(0.3)}$	92.8$_{(0.2)}$	**45.6**$_{(0.6)}$	52.3$_{(0.1)}$
1,600	45.2$_{(0.7)}$	**73.2**$_{(0.8)}$	92.9$_{(0.2)}$	**45.7**$_{(1.0)}$	51.5$_{(0.1)}$	43.7$_{(0.4)}$	**72.7**$_{(0.4)}$	92.7$_{(0.1)}$	**45.9**$_{(0.3)}$	51.0$_{(0.3)}$
3,200	45.2$_{(0.3)}$	72.3$_{(0.2)}$	92.8$_{(0.1)}$	44.3$_{(0.2)}$	50.9$_{(0.2)}$	44.8$_{(0.2)}$	72.2$_{(0.4)}$	92.9$_{(0.2)}$	45.3$_{(1.2)}$	50.4$_{(0.1)}$

Each score represents the average score of all experiments within that task group, encompassing different tasks, datasets, classification heads, and random seeds. The values colored with are higher than the Per-AA method. *Abbreviations,* **Vocab.**: vocabulary size; **Fit**: protein fitness; **PPI**: protein-protein interaction; **Loc**: protein localization; **Sol**: protein solubility; **Fold**: protein fold

† The top three are highlighted by First, Second, Third.

Table 8 The average results of different downstream task groups under the same vocabulary with varying pooling heads are presented

Vocab.	Mean Pooling					Attention1d Pooling				
	Fit	PPI	Loc	Sol	Fold	Fit	PPI	Loc	Sol	Fold
20	**41.5**$_{(0.2)}$	70.1$_{(0.7)}$	**92.8**$_{(0.2)}$	40.7$_{(1.2)}$	**56.8**$_{(0.2)}$	52.2$_{(0.3)}$	72.8$_{(0.7)}$	93.4$_{(0.1)}$	45.0$_{(1.5)}$	**57.2**$_{(0.2)}$
50	**41.1**$_{(0.2)}$	**70.6**$_{(0.3)}$	**93.0**$_{(0.2)}$	41.8$_{(0.1)}$	**56.2**$_{(0.2)}$	**53.7**$_{(0.5)}$	74.8$_{(0.4)}$	**93.7**$_{(0.1)}$	**49.0**$_{(0.6)}$	**57.2**$_{(0.5)}$
100	**41.5**$_{(0.2)}$	**70.1**$_{(0.3)}$	92.4$_{(0.0)}$	42.0$_{(0.1)}$	52.6$_{(0.4)}$	**53.1**$_{(0.2)}$	74.2$_{(0.1)}$	**93.8**$_{(0.2)}$	**48.4**$_{(0.7)}$	52.9$_{(0.2)}$
200	**41.2**$_{(0.3)}$	**70.8**$_{(0.3)}$	**92.5**$_{(0.1)}$	**42.7**$_{(0.7)}$	**54.3**$_{(1.0)}$	**52.4**$_{(0.3)}$	75.3$_{(0.3)}$	**93.7**$_{(0.1)}$	47.6$_{(0.9)}$	**53.6**$_{(0.4)}$
800	39.3$_{(0.4)}$	70.0$_{(0.2)}$	92.2$_{(0.2)}$	**42.9**$_{(0.7)}$	52.0$_{(0.4)}$	50.1$_{(0.3)}$	**75.5**$_{(0.1)}$	**93.6**$_{(0.1)}$	48.3$_{(1.5)}$	52.6$_{(0.4)}$
1,600	39.2$_{(0.3)}$	69.7$_{(0.2)}$	92.0$_{(0.1)}$	**43.2**$_{(0.5)}$	51.4$_{(0.4)}$	49.7$_{(0.6)}$	**76.1**$_{(0.6)}$	93.6$_{(0.2)}$	**48.4**$_{(0.8)}$	51.2$_{(0.5)}$
3,200	39.8$_{(0.1)}$	68.7$_{(0.6)}$	92.1$_{(0.1)}$	41.5$_{(0.3)}$	50.6$_{(0.1)}$	50.2$_{(0.1)}$	**75.7**$_{(0.2)}$	**93.6**$_{(0.1)}$	48.1$_{(0.7)}$	50.7$_{(0.1)}$

Each score represents the average score of all experiments within that task group, encompassing different tasks, datasets, classification heads, and random seeds. The values colored with are higher than the Per-AA method. Vocab.: vocabulary size; Fit: protein fitness; PPI: protein-protein interaction; Loc: protein localization; Sol: protein solubility; Fold: protein fold

† The top three are highlighted by First, Second, Third.

Discussion: Deeper Analysis

- **Minimal Impact of Tokenizer Method:**
  - The **choice between BPE and Unigram** had **minimal impact** on downstream task performance.
  - The **vocabulary size** itself is the dominant factor influencing model representation quality.
- **Impact of Pooling Heads:**
  - Performance is strongly dependent on the **selection of the classification head**.
  - **Attention1d Pooling** generally outperforms Mean Pooling.
  - *As vocabulary size increases, the representation capacity for Attention1d also tends to decrease (similar bell-shaped trend seen for tokenizers).*

Conclusion & Implications

- **Key Finding:** Expanding vocabulary size (to **50-200 elements**) generally **boosts PLM performance** on downstream tasks.
- **Critical Threshold:** Exceeding **~800 elements** often leads to a **significant drop** in model representation power.
- **PETA's Role:** Introduced a comprehensive benchmark for systematic PLM evaluation, setting a standard for future research.
- **Future Impact:** This work aims to influence the future PLM community, contributing positively to human health, environmental development, and biomedicine.
- **Open Resources:** Code, model weights, and datasets are publicly available on GitHub: https://github.com/ginnm/ProteinPretraining

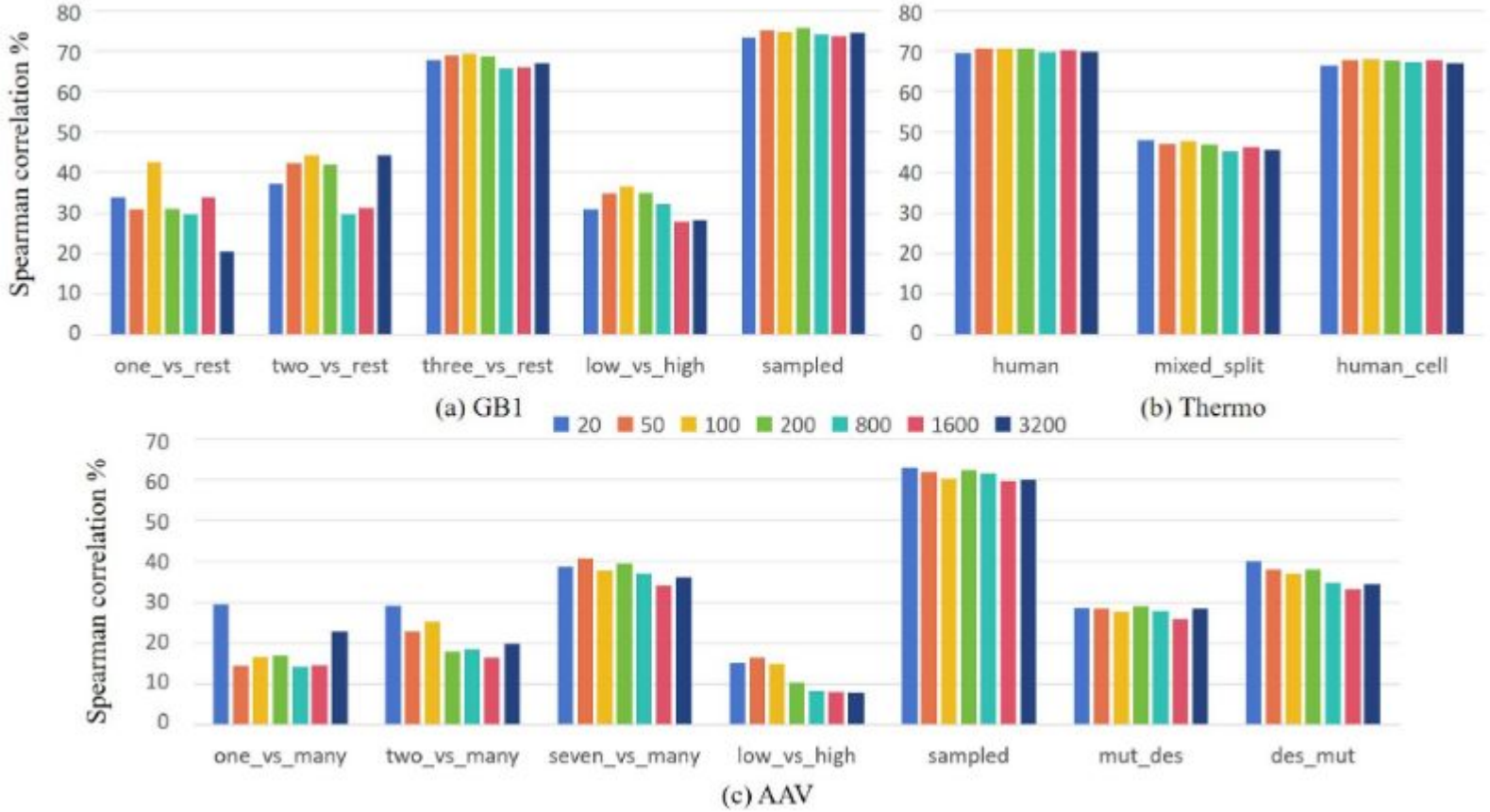


Fig. 3 Detail performances of the GB1, Thermo, and AAV datasets across different vocabulary sizes

Teşekkürler

The End.