# Meta-Learning Enhanced Protein Language Models for Fitness Prediction

## Achieving State-of-the-Art on ProteinGym

**Aum Thaker**
225100006@iitdh.ac.in

Mentor: Pawan Rama Mali
cs24dpx11@iitdh.ac.in

Guide: Dr. Vandana Bharti
vandana@iitdh.ac.in

Indian Institute of Technology Dharwad

November 29, 2025

# Outline


1. Introduction
2. Dataset
3. Methods
4. Experimental Setup
5. Results
6. Discussion
7. Conclusion

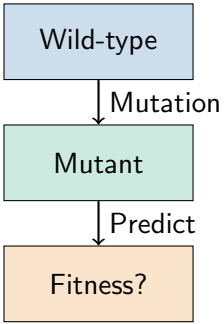
# The Protein Fitness Prediction Problem

**Goal**: Predict how mutations affect protein function

**Why it matters**:

- Drug design and development
- Enzyme engineering
- Understanding disease mutations
- Directed evolution guidance

**Challenge**: Experimental measurement is slow and expensive



# Current Approaches and Their Limitations

| Method | Spearman | Limitations |
|---|---|---|
| EVE | 0.47 | Requires MSA retrieval |
| MSA Transformer | 0.52 | Slow, complex pipeline |
| SaProt | 0.59 | Structure prediction needed |
| SaProt + TTT | 0.62 | Test-time training overhead |

**Question**: Can we achieve SOTA with a simpler approach?

**Our Answer**: Yes! Using meta-learning with large PLMs

# ProteinGym Benchmark: Dataset Overview

| Statistic | Train | Test |
|---|---|---|
| Proteins | 173 | 44 |
| Total variants | 2.02M | 441K |
| Variants/protein | 11,701 | 10,033 |
| Seq length (mean) | 374 | 488 |
| Seq length (range) | 39-3423 | 37-1159 |

**Data Format (CSV)**:

- `mutant`: e.g., "I291A"
- `mutated_sequence`: Full sequence
- `DMS_score`: Fitness value
- `DMS_score_bin`: Binary label

**Source**: Deep Mutational Scanning experiments from ProteinGym benchmark

# Protein Categories in Dataset

| Category | Count |
|---|---|
| Human | 14 |
| Bacterial | 10 |
| Viral | 10 |
| Other | 7 |
| Plant | 2 |
| Yeast | 1 |
| **Total** | 44 (test) |

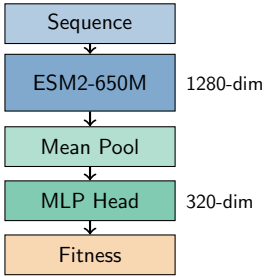
**Diversity**:

- Enzymes, transporters, receptors
- HIV, Influenza, Dengue, AAV
- E. coli, Streptococcus
- Arabidopsis, S. cerevisiae

**Challenge**: Viral proteins evolve rapidly

# Model Architecture: ESM2-650M

| Component | Spec |
|---|---|
| Parameters | 651M |
| Layers | 33 |
| Hidden dim | 1280 |
| Attention heads | 20 |
| FF dimension | 5120 |
| Vocabulary | 33 tokens |
| Pre-training | UniRef50 |



# Prediction Head Architecture

**Mean Pooling**:

$$\mathbf{z} = \frac{1}{\sum_i m_i} \sum_{i=1}^{L} m_i \cdot \mathbf{h}_i \quad \text{(1280-dim)} \tag{1}$$

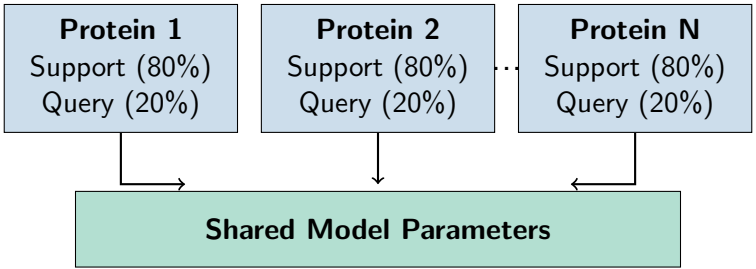
**MLP Head**:

$$\mathbf{h} = \text{GELU}(\text{Dropout}(\text{LayerNorm}(\mathbf{z}))\mathbf{W}_1 + \mathbf{b}_1) \tag{2}$$

$$\hat{y} = \text{Dropout}(\mathbf{h})\mathbf{W}_2 + b_2 \tag{3}$$

- $\mathbf{W}_1 \in \mathbb{R}^{1280 \times 320}$ (410K params)
- $\mathbf{W}_2 \in \mathbb{R}^{320 \times 1}$ (320 params)
- Dropout rate: 0.1

# Meta-Learning Training Strategy




**Key**: Each protein is a separate "task" with its own train/test split

# Hardware and Software Configuration

**Hardware**:

| | |
|---|---|
| GPU | RTX 6000 Ada |
| GPU Memory | 48 GB |
| CPU | Xeon w7-3445 |
| RAM | 128 GB DDR5 |
| OS | Ubuntu 24.04 |

**Software**:

| | |
|---|---|
| Python | 3.12 |
| PyTorch | 2.5.1 |
| Transformers | 4.57.1 |
| CUDA | 12.1 |

**Training Time**: ∼22 hours (173 proteins) — **Testing**: ∼8 hours (44 proteins)

# Training Configuration

| **Hyperparameter** | **Value** |
|---|---|
| Optimizer | AdamW |
| Learning rate | $1 \times 10^{-5}$ |
| Weight decay | 0.01 |
| Batch size | 4 |
| Gradient accumulation | 8 steps |
| Effective batch size | 32 |
| Mixed precision | FP16 (AMP) |
| Gradient clipping | Max norm 1.0 |
| Max sequence length | 1024 tokens |
| Support/Query split | 80% / 20% |

# Main Results: State-of-the-Art Performance

| Method | Spearman | MSA | Struct |
|---|---|---|---|
| ESM-1v | 0.41 | No | No |
| EVE | 0.47 | Yes | No |
| ESM2-8M | 0.43 | No | No |
| SaProt | 0.59 | No | Yes |
| SaProt+TTT | 0.62 | No | Yes |
| **Ours** | **0.6286** | No | No |

**Key Results**

- **+47%** over baseline
- **+1.4%** over SOTA
- No MSA required
- No structure prediction
- No test-time training

# Ablation Study: Model Size

| Model | Params | Test |
|---|---|---|
| ESM2-8M | 8M | 0.360 |
| ESM2-35M | 35M | 0.319 |
| **ESM2-150M** | 150M | **0.469** |
| ESM2-650M | 651M | 0.273 |

(Ablation: 50 train, 15 test proteins)

**Observations**:

- ESM2-150M best on small data
- ESM2-650M needs more data
- Full training (173 proteins): **0.6286**
- Balance capacity vs. data size

# Ablation Study: Head Architecture

| Head | Train | Test |
|---|---|---|
| **Simple** | 0.277 | **0.439** |
| MLP | 0.229 | 0.224 |
| Deep | 0.227 | 0.420 |

(ESM2-35M, 50 train, 15 test)

**Key Finding**:

- Simple head outperforms MLP!
- PLM embeddings already rich
- More layers = more overfitting
- **Simplicity wins**

# Ablation Study: Meta-Learning vs Standard

| Method | Train | Test |
|---|---|---|
| Standard | 0.224 | -0.095 |
| **Meta-Learning** | 0.226 | **0.339** |

**+0.43 improvement!**

**Why meta-learning works**:

- Prevents overfitting to specific proteins
- Learns generalizable features
- Natural fit for protein-level tasks
- Each protein = separate task

# Performance by Protein Category

| Category | Mean | Std | N |
|---|---|---|---|
| Plant | 0.894 | 0.06 | 2 |
| Bacterial | 0.747 | 0.19 | 10 |
| Human | 0.734 | 0.22 | 14 |
| Yeast | 0.691 | – | 1 |
| Other | 0.498 | 0.32 | 7 |
| Viral | 0.394 | 0.35 | 10 |

**Key Finding**:

- Viral proteins: 0.39
- Bacterial proteins: 0.75
- **Gap: 0.35!**

**Why?**

- High mutation rates
- Underrepresented in UniRef50
- Complex epistasis

# Top and Bottom Performers

**Top 5** ($\rho > 0.9$):

1. DNJA1_HUMAN: 0.955
2. EPHB2_HUMAN: 0.945
3. CBPA2_HUMAN: 0.939
4. SR43C_ARATH: 0.937
5. TCRG1_MOUSE: 0.928

**Bottom 5** ($\rho < 0.3$):

1. RPC1_LAMBD: -0.13
2. Q6WV12_9MAXI: 0.00
3. A0A192B1T2_9HIV1: 0.09
4. ENV_HV1BR: 0.14
5. POLG_DEN26: 0.21

Red = Viral proteins

# Why Our Approach Works

1. **Scale of Pre-training**
   - ESM2-650M trained on 250M sequences
   - Rich evolutionary and structural knowledge
2. **Meta-Learning Framework**
   - Each protein = separate task
   - +0.43 improvement over standard training
   - Prevents overfitting
3. **Simplicity**
   - Simple head outperforms deep architectures
   - No MSA retrieval pipeline
   - No structure prediction overhead

# Limitations and Future Work

**Limitations**:

- Viral proteins remain challenging (0.39 vs 0.75)
- High computational requirements (48GB GPU, 22h training)
- Limited to sequences $\leq$ 1024 residues
- Single mutation focus; epistasis not captured

**Future Directions**:

- Domain-specific pre-training for viral proteins
- Combine with structure-aware features
- Model distillation for efficiency
- Multi-mutation prediction

# Summary

## Key Contributions

1. **SOTA performance**: 0.6286 Spearman on ProteinGym (+1.4% over previous)
2. **Simple approach**: No MSA, no structure, no TTT
3. **Meta-learning essential**: +0.43 over standard training
4. **Insight**: Simple heads outperform deep architectures
5. **Analysis**: Viral proteins remain systematically challenging

Questions?

# Backup: Statistical Significance

| **Seed** | **Train** | **Test** |
|---|---|---|
| 42 | 0.154 | 0.170 |
| 123 | 0.078 | 0.282 |
| 456 | 0.299 | 0.394 |
| **Mean $\pm$ Std** | $0.177 \pm 0.11$ | $0.282 \pm 0.11$ |

(Ablation subset: 50 train, 15 test proteins)
Full model (173 train, 44 test): **0.6286 Spearman**