Novelty Assessment Report

Paper: Reverse Distillation: Disentangling and Scaling Protein Language Model Representations

PDF URL: https://openreview.net/pdf?id=f12Lo7ZUX5

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-04

Abstract

Unlike the foundation model scaling laws seen in natural language processing and computer vision, biological foundation models scale relatively poorly. For example, the ESM-2 family of protein language models plateaus at 650M-3B parameters on ProteinGym benchmarks. We address this limitation by introducing Reverse Distillation, a principled framework that decomposes large protein language model representations into orthogonal subspaces guided by smaller models of the same family. We hypothesize that this decomposition matches the natural hierarchy of protein properties, where broad features like secondary structure are robustly captured by compact, smaller models while the residual capacity of larger models specializes in protein-family specific functions. Our method is theoretically grounded and enables monotonic scaling---larger reverse-distilled models consistently outperform their smaller counterparts, overcoming the scaling plateau. Moreover, on ProteinGym benchmarks, reverse-distilled ESM-2 variants broadly outperform their respective baseline models at the same embedding dimensionality. Our approach offers a generalizable framework for disentangling hierarchical feature spaces in foundation model embeddings, with potential applications across biology and other domains where scaling challenges persist.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Improving Scaling Behavior of Protein Language Models Through Representation Decomposition**

A total of **1 papers** were analyzed and organized into a taxonomy with **2 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Representation Decomposition and Distillation Methods**
- **Retrieval-Based Augmentation for Representation Learning**

Complete Taxonomy Tree

- Improving Scaling Behavior of Protein Language Models Through Representation Decomposition Survey Taxonomy
- Representation Decomposition and Distillation Methods
  - Hierarchical Feature Disentanglement via Reverse Distillation ★ (1 papers)
  - [0] Reverse Distillation: Disentangling and Scaling Protein Language Model Representations (Anon et al., 2026) View paper
- Retrieval-Based Augmentation for Representation Learning
  - Retrieved Sequence Augmentation (1 papers)
  - [1] Retrieved sequence augmentation for protein representation learning (Chang Ma, 2024) View paper

Narrative

Core task: Improving scaling behavior of protein language models through representation decomposition. The field addresses how to enhance protein language models by breaking down their learned representations into more interpretable or efficient components. The taxonomy reveals two main branches: one focused on Representation Decomposition and Distillation Methods, which explores techniques for disentangling hierarchical features and compressing knowledge from larger models, and another centered on Retrieval-Based Augmentation for Representation Learning, which leverages external sequence databases to enrich model representations. These branches reflect complementary strategies—internal restructuring of learned features versus external augmentation through retrieved information—both aiming to improve model scalability and performance on protein-related tasks.

Within the decomposition branch, works like Reverse Distillation[0] pursue hierarchical feature disentanglement by reversing traditional distillation flows, aiming to isolate different levels of structural or functional information encoded in protein representations. This contrasts with retrieval-based approaches such as Retrieved Sequence Augmentation[1], which augment model inputs or embeddings by incorporating similar sequences from large databases, thereby grounding predictions in evolutionary context. The original paper sits squarely within the decomposition paradigm, emphasizing how reverse distillation can untangle complex representations to achieve better scaling properties. Compared to retrieval methods that rely on external data, Reverse Distillation[0] focuses on intrinsic model architecture and training dynamics, offering a complementary path toward more efficient and interpretable protein language models as datasets and model sizes continue to grow.

Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

Contributions Analysis

Overall novelty summary. The paper introduces Reverse Distillation, a framework for decomposing protein language model representations into orthogonal subspaces guided by smaller models. Within the taxonomy, it occupies the sole position in the 'Hierarchical Feature Disentanglement via Reverse Distillation' leaf under 'Representation Decomposition and Distillation Methods'. This leaf contains only the original paper itself, indicating a sparse research direction with no sibling papers identified in the taxonomy structure. The approach targets the scaling plateau observed in ESM-2 models, aiming to enable monotonic performance improvements as model size increases.

The taxonomy reveals two main branches: representation decomposition methods and retrieval-based augmentation approaches. The original paper sits within the decomposition branch, which focuses on internal restructuring of learned features rather than external data augmentation. The neighboring 'Retrieved Sequence Augmentation' leaf represents an alternative strategy that enhances representations through database retrieval. The taxonomy's scope notes clarify that methods using model-guided decomposition belong in the original paper's branch, while those relying on external sequence databases fall into the retrieval category, suggesting these represent distinct methodological paradigms within protein language model research.

Among 25 candidates examined across three contributions, no clearly refutable prior work was identified. The core Reverse Distillation framework examined 10 candidates with zero refutations, the hierarchical decomposition with theoretical guarantees examined 10 candidates with zero refutations, and the Matryoshka-style embeddings examined 5 candidates with zero refutations. This limited search scope suggests that within the top-25 semantically similar papers, no direct overlaps were detected. However, the analysis explicitly notes this is not an exhaustive literature review, and the sparse taxonomy structure (only one paper in the leaf) may reflect either genuine novelty or limitations in the search methodology.

Based on the limited search of 25 candidates, the work appears to occupy a relatively unexplored niche within protein language model scaling. The absence of sibling papers in the taxonomy and zero refutable candidates across all contributions suggest potential novelty, though this assessment is constrained by the search scope. The taxonomy structure indicates the field has alternative approaches (retrieval-based methods) but limited prior work specifically on reverse distillation for hierarchical feature decomposition in protein models.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Reverse Distillation framework for decomposing protein language model representations

Description: The authors propose a method that decomposes large protein language model representations into orthogonal subspaces guided by smaller models from the same family. This decomposition separates universal features (captured by smaller models) from specialized features (unique to larger models), addressing the scaling plateau observed in biological foundation models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. PREreview of "InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders"

URL: View paper

Brief Assessment

InterPLM Review[13] focuses on sparse autoencoders (SAEs) for extracting interpretable features from protein language models, not on decomposing representations into orthogonal subspaces guided by smaller models from the same family.

---

2. InterPLM: discovering interpretable features in protein language models via sparse autoencoders.

URL: View paper

Brief Assessment

InterPLM[12] focuses on discovering interpretable features in protein language models using sparse autoencoders, not on decomposing representations into orthogonal subspaces guided by smaller models from the same family.

---

3. ESM-MHC: An Improved Predictor of MHC Using ESM Protein Language Model

URL: View paper

Brief Assessment

ESM-MHC[8] uses ESM protein language models only for feature extraction followed by PCA dimensionality reduction for MHC prediction. It does not propose any method for decomposing or disentangling protein language model representations into orthogonal subspaces.

---

4. Mechanistic Interpretability of Fine-Tuned Protein Language Models for Nanobody Thermostability Prediction

URL: View paper

Brief Assessment

Nanobody Thermostability[11] focuses on mechanistic interpretability of fine-tuned protein language models for thermostability prediction using sparse representations, not on decomposing representations into orthogonal subspaces guided by smaller models from the same family.

---

5. PLM-eXplain: Divide and Conquer the Protein Embedding Space

URL: View paper

Brief Assessment

PLM-eXplain[14] decomposes embeddings into interpretable biochemical features (secondary structure, hydropathy) versus residual components for classification tasks, not into hierarchical universal/specialized subspaces guided by smaller models from the same family to address scaling plateaus.

---

6. Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction

URL: View paper

Brief Assessment

ESM2 Feature Extraction[7] focuses on extracting and analyzing different types of features from ESM2 for subcellular localization prediction, not on decomposing representations into orthogonal subspaces guided by smaller models from the same family.

---

7. Dynamic insights into the structural evolution of ACE2â□□RBD interactions through molecular dynamics simulation, Markov state modeling, and large language model â□¦

URL: View paper

Brief Assessment

ACE2-RBD Dynamics[16] focuses on molecular dynamics simulation and Markov state modeling of protein-protein interactions, not on decomposing protein language model representations into interpretable subspaces.

---

8. Sparse autoencoders uncover biologically interpretable features in protein language model representations

URL: View paper

Brief Assessment

Sparse Autoencoders Proteins[10] focuses on extracting interpretable features from protein language models using sparse autoencoders and transcoders, not on decomposing representations into orthogonal subspaces guided by smaller models from the same family.

9. ProtSAE: Disentangling and Interpreting Protein Language Models via Semantically-Guided Sparse Autoencoders

URL: View paper

Brief Assessment

ProtSAE[9] focuses on semantic disentanglement within a single protein language model using sparse autoencoders and biological annotations, not on decomposing representations across multiple model scales or addressing scaling plateaus through orthogonal subspace decomposition.

10. Sparse Autoencoders for Low-N Protein Function Prediction and Design

URL: View paper

Brief Assessment

Low-N Protein Design[15] focuses on sparse autoencoders (SAEs) for decomposing fine-tuned ESM2 embeddings into interpretable latent variables for low-N protein function prediction and design tasks. The original paper proposes reverse distillation to decompose large PLM representations into orthogonal subspaces guided by smaller models from the same family to address scaling plateaus. These are fundamentally different decomposition approaches with different objectives and methodologies.

Contribution 2: Hierarchical decomposition with theoretical optimality guarantees

Description: The method provides a theoretically grounded hierarchical decomposition where each model scale contributes orthogonal information. The authors prove this decomposition is MSE-optimal among all representations that preserve the smaller model's embeddings, ensuring quality approximation of the original space.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Computational modeling of hierarchically polarized groups by structured matrix factorization

URL: View paper

Brief Assessment

Hierarchical Polarized Groups[21] addresses hierarchical decomposition of social media belief structures using matrix factorization, not protein language model embeddings. The domains, methods, and theoretical guarantees are fundamentally different.

2. Hierarchical Approximate Proper Orthogonal Decomposition

URL: View paper

Brief Assessment

Hierarchical POD[26] focuses on computational efficiency for POD approximation in numerical simulations, not on disentangling protein language model representations or addressing scaling plateaus in biological foundation models.

3. Bi-Level Orthogonal Multi-Teacher Distillation

URL: View paper

Brief Assessment

Bi-Level Orthogonal Distillation[25] focuses on multi-teacher knowledge distillation for image classification tasks using orthogonal projections to align teacher-student feature spaces, not on hierarchical decomposition of language model embeddings with MSE-optimality guarantees for preserving smaller model representations.

4. Multi-level attention-based domain disentanglement for BCDR

URL: View paper

Brief Assessment

Multi-level Domain Disentanglement[20] focuses on cross-domain recommendation systems using domain-invariant and domain-specific features for user preferences, not hierarchical decomposition of language model embeddings with MSE-optimality guarantees.

5. A Bayesian Hierarchical Model for Orthogonal Tucker Decomposition with Oblivious Tensor Compression

URL: View paper

Brief Assessment

The candidate paper's full text is not available (marked as 'n/a'), making it impossible to assess whether it refutes the novelty of the original paper's hierarchical decomposition with MSE-optimality guarantees for protein language model representations.

6. OrtSAE: Orthogonal Sparse Autoencoders Uncover Atomic Features

URL: View paper

Brief Assessment

OrtSAE[23] focuses on enforcing orthogonality between sparse autoencoder features to address feature absorption and composition in interpretability research, not on hierarchical decomposition of model embeddings with MSE-optimality guarantees for multi-scale protein language models.

7. Ethos: Rectifying Language Models in Orthogonal Parameter Space

URL: View paper

Brief Assessment

Ethos[18] focuses on task arithmetic for rectifying language models (debiasing, detoxification, memorization unlearning) using SVD to identify undesired knowledge components. It does not address hierarchical decomposition of model embeddings across different scales or provide MSE-optimality guarantees for preserving smaller model representations.

8. Two Heads are Better than One: Distilling Large Language Model Features Into Small Models with Feature Decomposition and Mixture

URL: View paper

Brief Assessment

Two Heads Distillation[17] focuses on market making tasks using LLM distillation with orthogonal feature decomposition across layer, task, and data dimensions. This is fundamentally different from the original paper's hierarchical decomposition of protein language model embeddings with MSE-optimality guarantees for preserving smaller model representations.

9. Latent symbol lattices in probabilistic semiosis: An unconventional architectural mechanism for contextual modulation in large language models

URL: View paper

Brief Assessment

Latent Symbol Lattices[19] focuses on contextual modulation mechanisms in LLMs using lattice structures, not on hierarchical decomposition of embeddings with MSE-optimality guarantees for protein language models.

10. TripleFDS: Triple Feature Disentanglement and Synthesis for Scene Text Editing

URL: View paper

Brief Assessment

TripleFDS[24] addresses scene text editing through feature disentanglement of style, content, and background attributes, which is fundamentally different from the original paper's hierarchical decomposition of protein language model embeddings with MSE-optimality guarantees.

Contribution 3: Matryoshka-style embeddings enabling monotonic scaling

Description: The framework produces embeddings with a nested prefix structure where smaller-dimensional prefixes correspond to valid reverse-distilled representations at that scale. This enables controlled performance degradation as embedding size decreases and restores monotonic scaling behavior where larger models consistently outperform smaller ones.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Visual data mining using monotone Boolean functions

URL: View paper

Brief Assessment

Visual Data Mining[5] focuses on visual data representation using monotone Boolean functions for data mining tasks, not on protein language model embeddings or nested prefix structures for controlled performance degradation.

2. Multi-layer logicâ□□A predicate logic including data structure as knowledge representation language

URL: View paper

Brief Assessment

Multi-layer Logic[4] is a predicate logic framework for knowledge representation, not a machine learning embedding method. It does not address neural network embeddings, scaling behavior, or representation learning.

3. Monotone Deep Spectrum Kernels

URL: View paper

Brief Assessment

The candidate paper context is too fragmentary to assess whether it refutes the original's novelty claim about nested prefix embeddings with monotonic scaling behavior in protein language models.

4. Latent surface alignment through oscillatory token drift in instruction-following large language models

URL: View paper

Brief Assessment

Oscillatory Token Drift[3] focuses on token-level dynamics in instruction-following LLMs through latent surface alignment, not on nested embedding structures or monotonic scaling behavior in protein language models.

5. Toward a Unified Theory of Time: Time as Monotonic Information Flow Across Causal Interfaces

URL: View paper

Brief Assessment

Unified Time Theory[2] discusses time as monotonic information flow and causal interfaces in physics, not machine learning embeddings or nested representation structures for computational models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Reverse Distillation: Disentangling and Scaling Protein Language Model Representations View paper
- [1] Retrieved sequence augmentation for protein representation learning View paper
- [2] Toward a Unified Theory of Time: Time as Monotonic Information Flow Across Causal Interfaces View paper
- [3] Latent surface alignment through oscillatory token drift in instruction-following large language models View paper
- [4] Multi-layer logicâ□□A predicate logic including data structure as knowledge representation language View paper
- [5] Visual data mining using monotone Boolean functions View paper
- [6] Monotone Deep Spectrum Kernels View paper
- [7] Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction View paper
- [8] ESM-MHC: An Improved Predictor of MHC Using ESM Protein Language Model View paper
- [9] ProtSAE: Disentangling and Interpreting Protein Language Models via Semantically-Guided Sparse Autoencoders View paper
- [10] Sparse autoencoders uncover biologically interpretable features in protein language model representations View paper
- [11] Mechanistic Interpretability of Fine-Tuned Protein Language Models for Nanobody Thermostability Prediction View paper
- [12] InterPLM: discovering interpretable features in protein language models via sparse autoencoders. View paper
- [13] PREreview of "InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders" View paper
- [14] PLM-eXplain: Divide and Conquer the Protein Embedding Space View paper

- [15] Sparse Autoencoders for Low-N Protein Function Prediction and Design View paper
- [16] Dynamic insights into the structural evolution of ACE2âRBD interactions through molecular dynamics simulation, Markov state modeling, and large language model â¦ View paper
- [17] Two Heads are Better than One: Distilling Large Language Model Features Into Small Models with Feature Decomposition and Mixture View paper
- [18] Ethos: Rectifying Language Models in Orthogonal Parameter Space View paper
- [19] Latent symbol lattices in probabilistic semiosis: An unconventional architectural mechanism for contextual modulation in large language models View paper
- [20] Multi-level attention-based domain disentanglement for BCDR View paper
- [21] Computational modeling of hierarchically polarized groups by structured matrix factorization View paper
- [22] A Bayesian Hierarchical Model for Orthogonal Tucker Decomposition with Oblivious Tensor Compression View paper
- [23] OrtSAE: Orthogonal Sparse Autoencoders Uncover Atomic Features View paper
- [24] TripleFDS: Triple Feature Disentanglement and Synthesis for Scene Text Editing View paper
- [25] Bi-Level Orthogonal Multi-Teacher Distillation View paper
- [26] Hierarchical Approximate Proper Orthogonal Decomposition View paper