Hierarchical Contrastive Learning for Enzyme Function Prediction

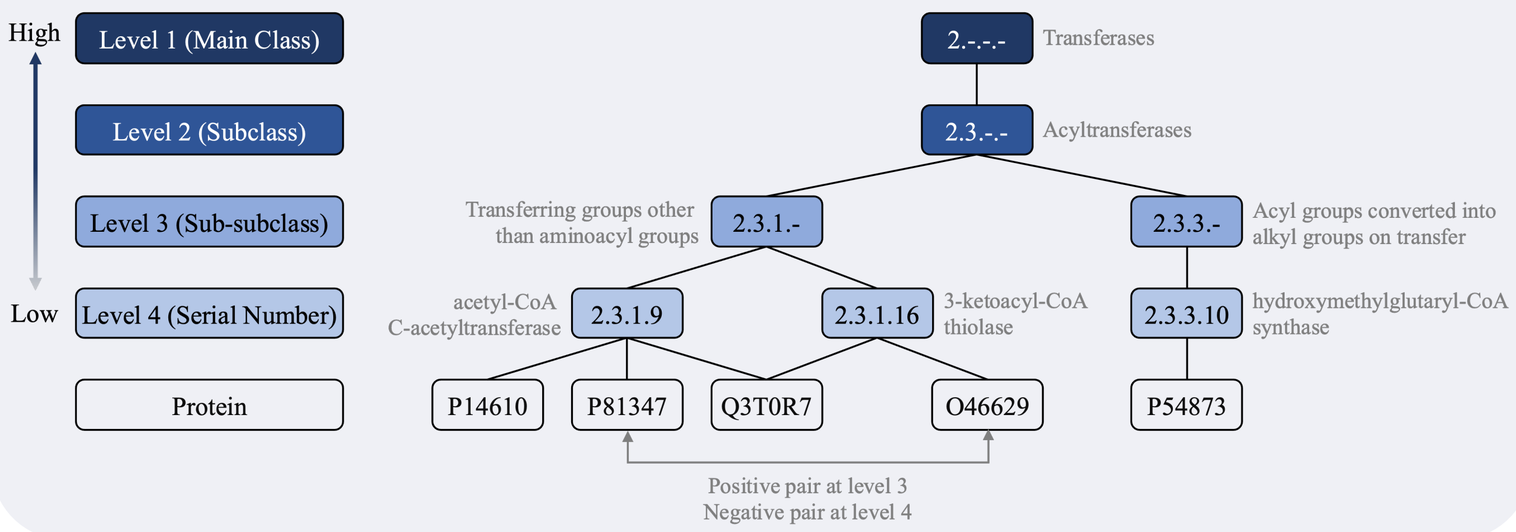
Soorin Yim, Doyeong Hwang, Kiyoung Kim, Sehui Han

LG AI Research

LG AI Research

INTRODUCTION

- **Background**: enzymes are biological catalysts with numerous industrial applications, and they are classified by the Enzyme Commission (EC) number system.
- **Motivation**: EC number prediction is challenged by class imbalance and the intrinsic hierarchy of the EC number system.
- We employed **hierarchical contrastive learning** to address these challenges.



METHODS

- We employed **Hierarchical Multi-label Contrastive (HMC) loss** (Zhang et al., 2022), which extends supervised contrastive loss across all hierarchical levels.
- Pair loss between an anchor i and a positive sample p at hierarchical level l:

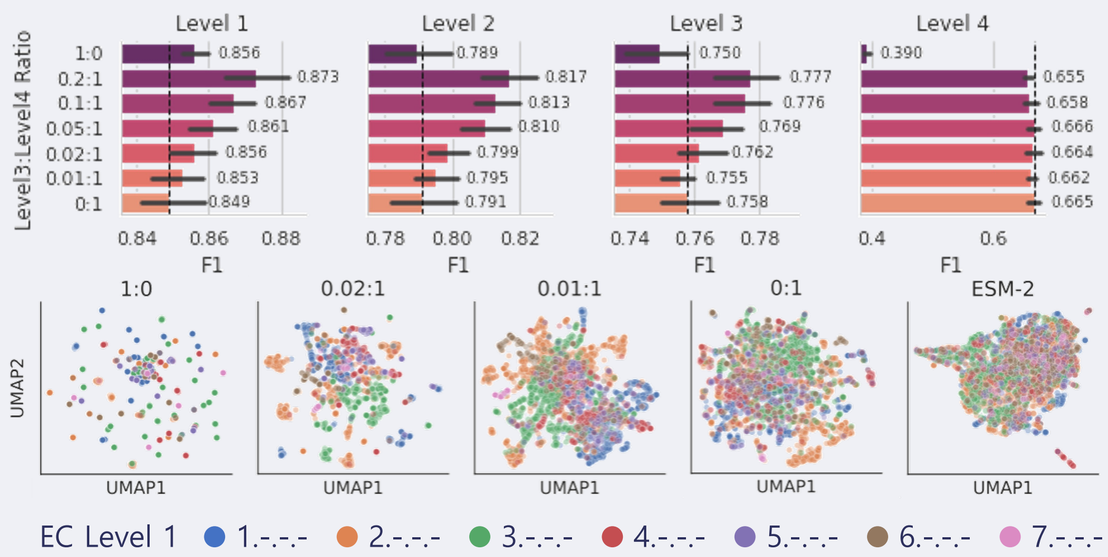
$$L^{\text{pair}} = \log \frac{\exp(f_i \cdot f_p^l / \tau)}{\sum_{a \in A_i} exp(f_i \cdot f_a / \tau)}$$

- **HMC loss:**

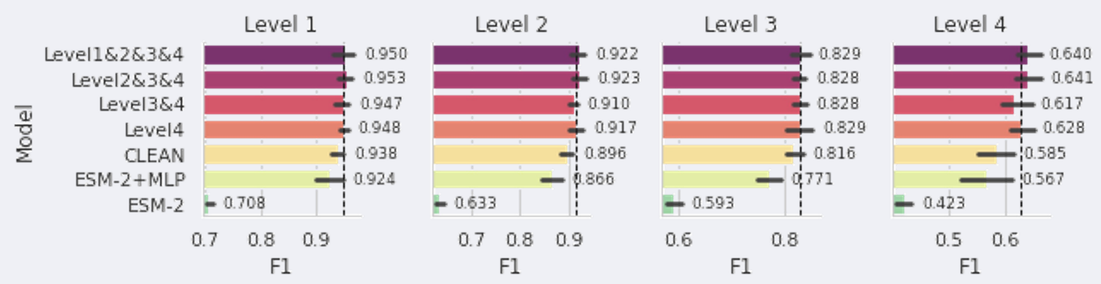
$$L^{\text{HMC}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P_l(i)|} \sum_{p_l \in P_l} L^{\text{pair}}(i, p_l^i)$$

RESULTS

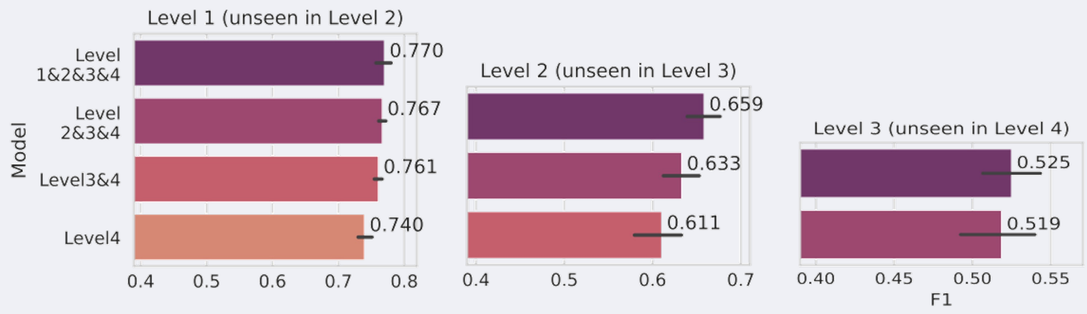
- **Datasets**
  - Split30 from Yu et al., 2023: 10,202 proteins, 3,576 EC numbers
  - New-392: 392 proteins, 177 EC numbers
  - **3 proteins per EC numbers** on average
- **Model architecture:** protein sequences were embedded with ESM-2-650M (Lin et al., 2023), followed by multi-layer perceptron (MLP) with three hidden layers and layer normalization.
- **HMC for level 3 and level 4**
  - Stronger weights to level 3 enforces EC number hierarchy while compromising performance for level 4 EC numbers.
  - Optimal weight enhances level 3 performance while preserving level 4 performance.



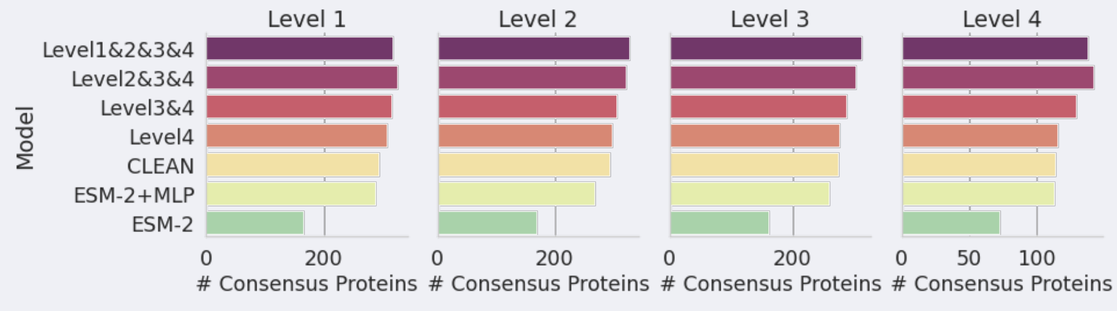
- **Extension of HMC to all levels**



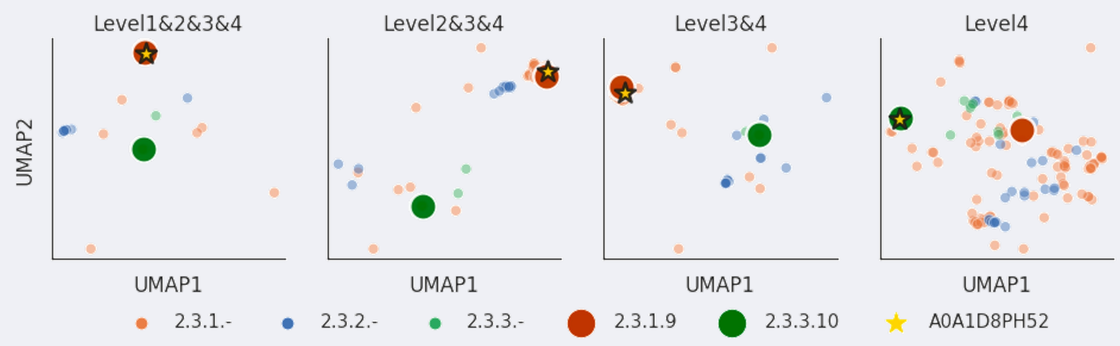
- **Prediction of unseen EC numbers**



- **Robustness**



- **Case study for identifying the function of a new enzyme**
  Initially misclassified as 2.3.3.10, incorporating EC number hierarchy correctly identifies A0A1D8PH52 as 2.3.1.9.



- **Conclusion**: hierarchical contrastive learning on EC number prediction improved
  - Higher-level performance with lowest-level performance retained.
  - Robustness.
  - Generalization on unseen EC numbers.