A Hierarchical Domain Adaptation Method in Neural Language Models

With Application to Taxonomy-Aware Linear B-cell Epitope Prediction

---

Lindeberg Pessoa Leite

PhD candidate

Date: August 28, 2025

Universidade de Brasilia (UnB)
Supervisor: Dr. Teófilo Emidio de Campos
Co-Supervisor: prof. Dr. Felipe Campelo

Overview

- Motivation
- Research hypothesis
- Related Work
- Proposed Method
- Case Study / Results
- Limitations
- Achievements
- Future Work

Motivation

- **Domain adaptation** is a central challenge in machine learning.
- In real scenarios, domain adaptation frequently involves source domains with **internal hierarchical structures** (e.g., phylogenetic branching, evolutionary progression of languages).
- **Multi-source domain adaptation** is a well-studied approach for this type of problem.

→ **However, hierarchical adaptation of the source domain remains a less researched area.**

Research Hypothesis

Context: Many real-world datasets are hierarchically organized, with abundant data at higher levels and scarce data at lower levels.

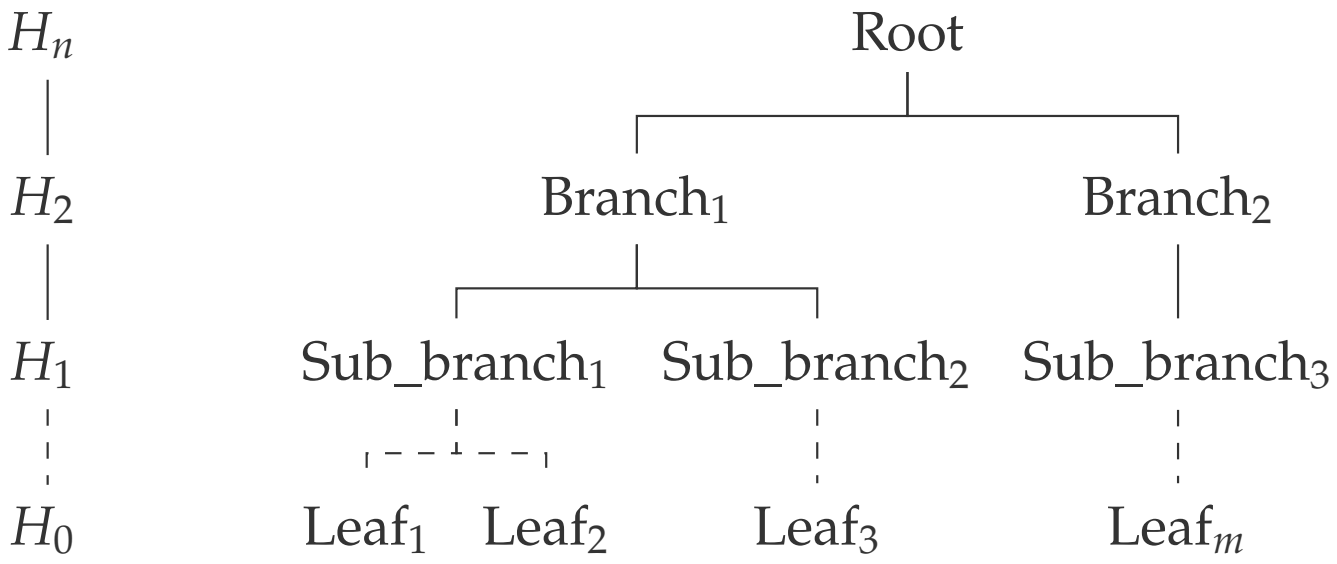
Hypothesis: *Given hierarchically structured data, how can knowledge be effectively transferred from data-rich higher levels to data-scarce lower levels?*

Related Work

Hierarchical Domain Adaptation

- Mind the Gap: Subspace based HDA [Raj et al., 2014]
- HDA with local feature patterns [Wen et al., 2022]
- Efficient HDA for PLM [Chronopoulou et al., 2022]

Proposed method – data structure



Hierarchical structure example: Branch$_1$ could represent the source domain (D_S), while Leaf$_3$ serves as the target domain (D_T).

Proposed method - total cost

The total cost, denoted as Weighted Cross Entropy Loss (WCEL), is defined as follows:

$$\mathrm{WCEL} = \min_{\theta} \Bigg[\underbrace{\sum_{l=0}^{n} \frac{w_l}{\sum_{j=0}^{n} w_j} \mathrm{CE}\Big(f_{\theta,l}\big(f_{\mathrm{map}}(x_s,\, H_{indexS})\big),\, y_s;\, \alpha_l \Big)}_{\text{Source-Domain Loss}} + \underbrace{\sum_{(x_t,\, y_t) \in (X_T,\, Y_T)} \mathrm{CE}\Big(f_{\theta,l}\big(f_{\mathrm{map}}(x_t,\, H_{indexT})\big),\, y_t;\, \alpha_l \Big)}_{\text{Target-Domain Loss}} \Bigg].$$

Case study: Epitope prediction

Hierarchical data in a phylogenetic structure.

Epitope identification is crucial for diagnostics (early disease detection), immunotherapy (personalised treatments), and immunisation (vaccine design).

ML and DL are now standard approaches.

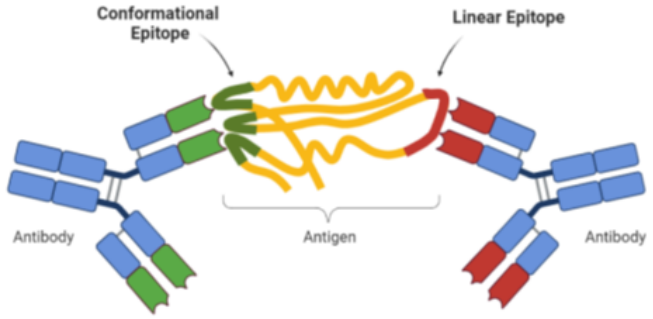


Image by Jodie Ashford/BioRender

Related Work

- Organism-Specific [Ashford et al., 2021]
- EpitopeVec [Bahai et al., 2021]
- EpiDope [Collatz et al., 2020]
- ESM-2 [Lin et al., 2023]
- BepiPred-3.0 [Clifford et al., 2022]

Single Domain Adaptation

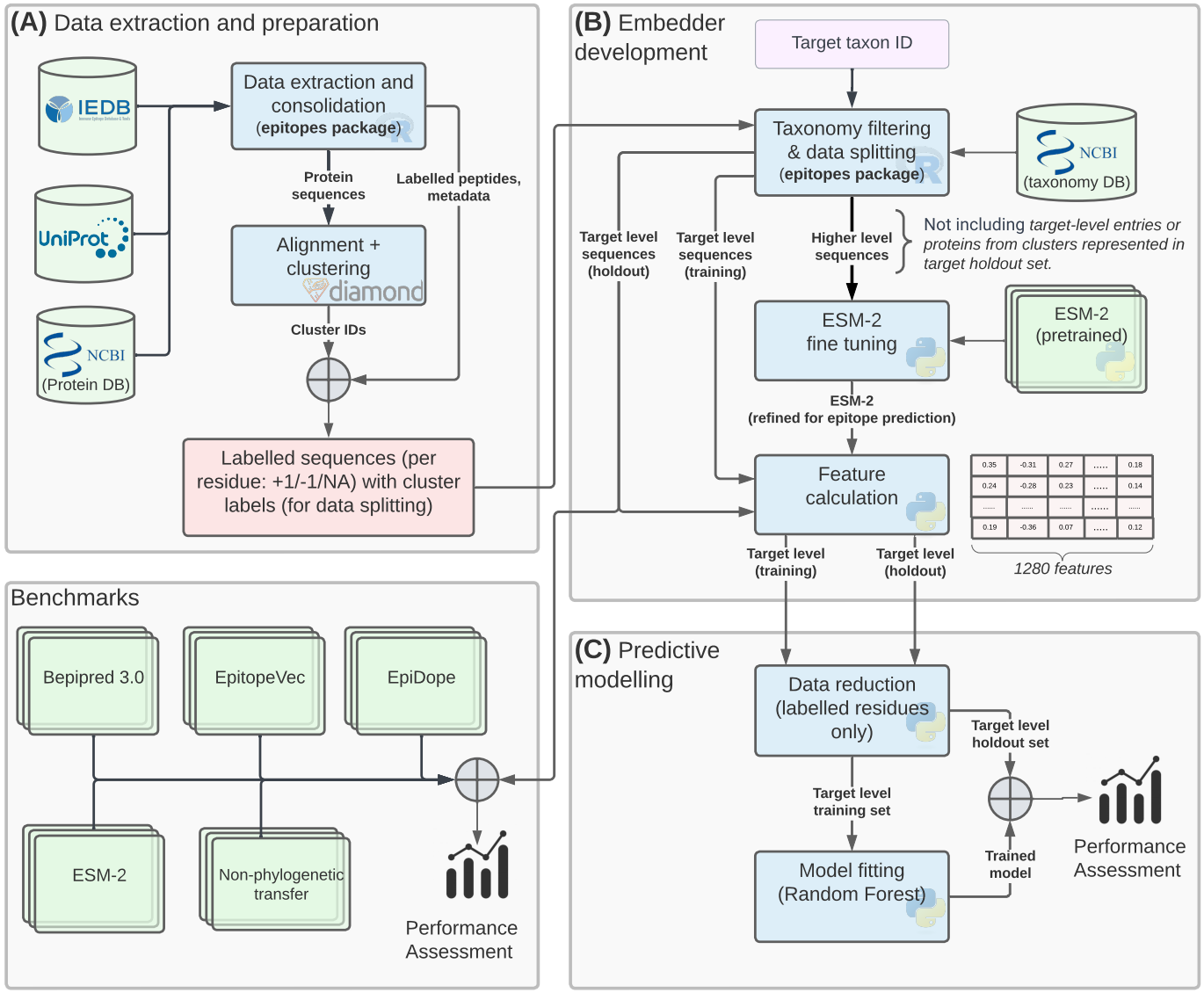


Figure: Submitted to *Genomics, Proteomics & Bioinformatics (Impact Factor: 11.5)* – under review

Results — Single Domain Adaptation (SDA)

Metrics: AUC, BACC, F1, MCC, NPV, PPV, Sensitivity, Specificity — Wilcoxon test with FDR correction

- 20 datasets; compared models: BepiPred 3, EpiDope, EpitopeVec, ESM-2, NPTransfer.
- EpitopeTransfer significantly outperformed most predictors on key metrics (AUC, BACC, F1, MCC), with some exceptions in PPV and Sensitivity
- **Filoviridae (MCC = 0.76) and Plasmodium falciparum (MCC = 0.50)** showed strong performance.
- No significant difference between ESM-1b and ESM-2.

Hierarchical Domain Adaptation

- **Datasets:** *B. pertussis*, *E. coli*, and *M. tuberculosis*
- **Trainable layers:** 4
- **Number of trials:** 5

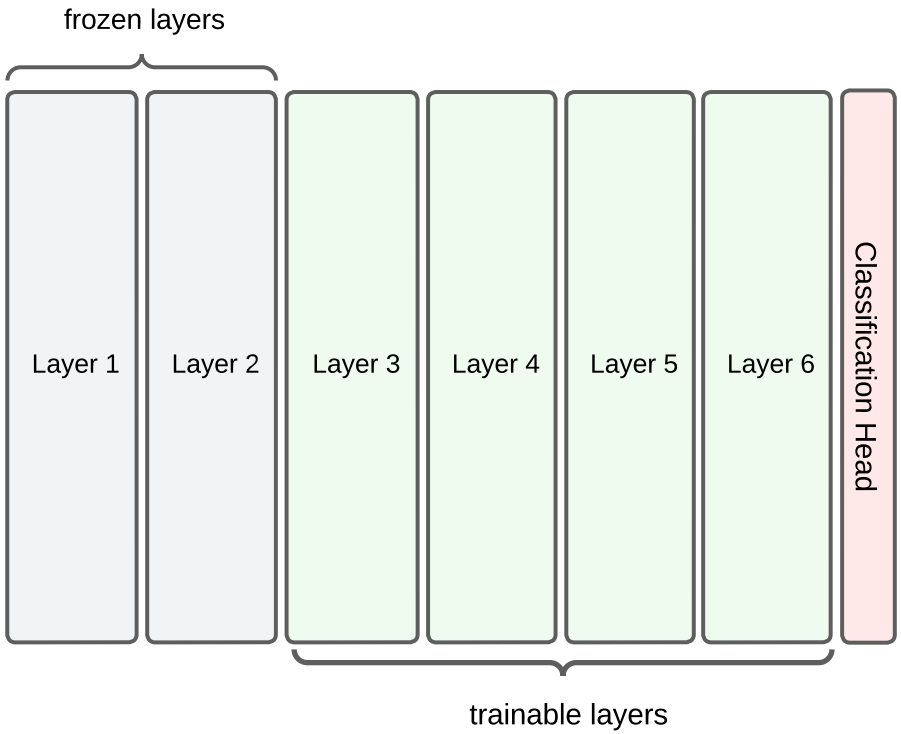


Figure: ESM-2 with 6 Transformer encoder layers and 8M parameters

Results — Hierarchical Domain Adaptation (HDA)

Metrics: AUC, BACC, F1, MCC, NPV, PPV, Sensitivity, Specificity — Wilcoxon test with FDR correction

- ESM2-8M
- 17 datasets; baseline ignores hierarchy.
- EpitopeTransfer significantly outperformed the baseline on most metrics except PPV and Specificity.
- **Filoviridae (MCC = 0.595) and Mononegavirales (MCC = 0.483)** showed strong performance.

Limitations

- **HDA-8M vs. SDA-8M Performance:** HDA-8M showed numerical gains over SDA-8M in almost all metrics, except Specificity. But only AUC gain was statistically significant ($p = 0.0109$, median diff = 0.039).
- **Data Constraint:** HDA relied on subsampled data due to resource limits → results may underrepresent full hierarchy.
- **SDA Weaknesses:** Underperformed for *SARS-CoV-2* (MCC=0.043 vs 0.169 baseline) and *M. tuberculosis* (MCC=-0.031 vs 0.039 baseline).

Achievements

- **LayoutQT** – Published in *Engineering Applications of Artificial Intelligence - 2023* (Qualis A1) [de Lucena Drumond et al., 2023].
- **EpitopeTransfer** – Submitted to *Genomics, Proteomics & Bioinformatics - 2024* (Impact Factor: 11.5) – under review. Preprint: [Leite et al., 2025].
- **Extended abstract** – Selected for oral presentation at the *International Conference on Intelligent Systems for Molecular Biology* [ISMB/ECCB, 2025].
- **Book Chapter** – Invited contribution to *Artificial Neural Networks, 4th ed.* (Methods in Molecular Biology series).
- **Hierarchical Domain Adaptation** – New article in preparation.

Future Work

- Improve optimization by refining stopping criteria to better prevent overfitting across taxa.
- Extend HDA to fully unsupervised scenarios without target domain labels.
- Add interpretability methods to highlight input regions influencing predictions.

References I

Jodie Ashford, João Reis-Cunha, Igor Lobo, Francisco Lobo, and Felipe Campelo. Organism-specific training improves performance of linear b-cell epitope prediction. *Bioinformatics*, 37(24):4826–4834, 07 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab536. URL https://doi.org/10.1093/bioinformatics/btab536.

Akash Bahai, Ehsaneddin Asgari, Mohammad R K Mofrad, Andreas Kloetgen, and Alice C McHardy. Epitopevec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics*, 37(23):4517–4525, 06 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab467. URL https://doi.org/10.1093/bioinformatics/btab467.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. Efficient hierarchical domain adaptation for pretrained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.96. URL https://aclanthology.org/2022.naacl-main.96/.

Joakim Nøddeskov Clifford, Magnus Haraldson Høie, Sebastian Deleuran, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *Protein Science*, 31 (12):e4497, 2022. doi: https://doi.org/10.1002/pro.4497. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4497.

Maximilian Collatz, Florian Mock, Emanuel Barth, Martin Hölzer, Konrad Sachse, and Manja Marz. Epidope: a deep neural network for linear b-cell epitope prediction. *Bioinformatics*, 37(4):448–455, 09 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa773. URL https://doi.org/10.1093/bioinformatics/btaa773.

Patricia Medyna Lauritzen de Lucena Drumond, Lindeberg Pessoa Leite, Teofilo E. de Campos, and Fabricio Ataides Braz. Layoutqt—layout quadrant tags to embed visual features for document analysis. *Engineering Applications of Artificial Intelligence*, 122:106091, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2023.106091. URL https://www.sciencedirect.com/science/article/pii/S0952197623002750.

References II

ISMB/ECCB, editor. *ISMB/ECCB 2025: Book of Abstracts*, 2025. International Society for Computational Biology (ISCB). URL https://www.iscb.org/images/stories/ismbeccb2025/document.AbstractsBook.ISMBECCB.2025.pdf. Proceedings of the Joint Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology.

Lindeberg Pessoa Leite, Teófilo Emidio de Campos, Francisco Pereira Lobo, and Felipe Campelo. Epitopetransfer: a phylogeny-aware transfer learning framework for taxon-specific linear b-cell epitope prediction. *bioRxiv*, 2025. doi: 10.1101/2025.04.17.649425. URL https://www.biorxiv.org/content/early/2025/06/27/2025.04.17.649425.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Anant Raj, Vinay P Namboodiri, and Tinne Tuytelaars. Mind the gap: Subspace based hierarchical domain adaptation, 2014. URL https://arxiv.org/abs/1501.03952.

Jun Wen, Junsong Yuan, Qian Zheng, Risheng Liu, Zhefeng Gong, and Nenggan Zheng. Hierarchical domain adaptation with local feature patterns. *Pattern Recognition*, 124:108445, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108445. URL https://www.sciencedirect.com/science/article/pii/S003132032100621X.