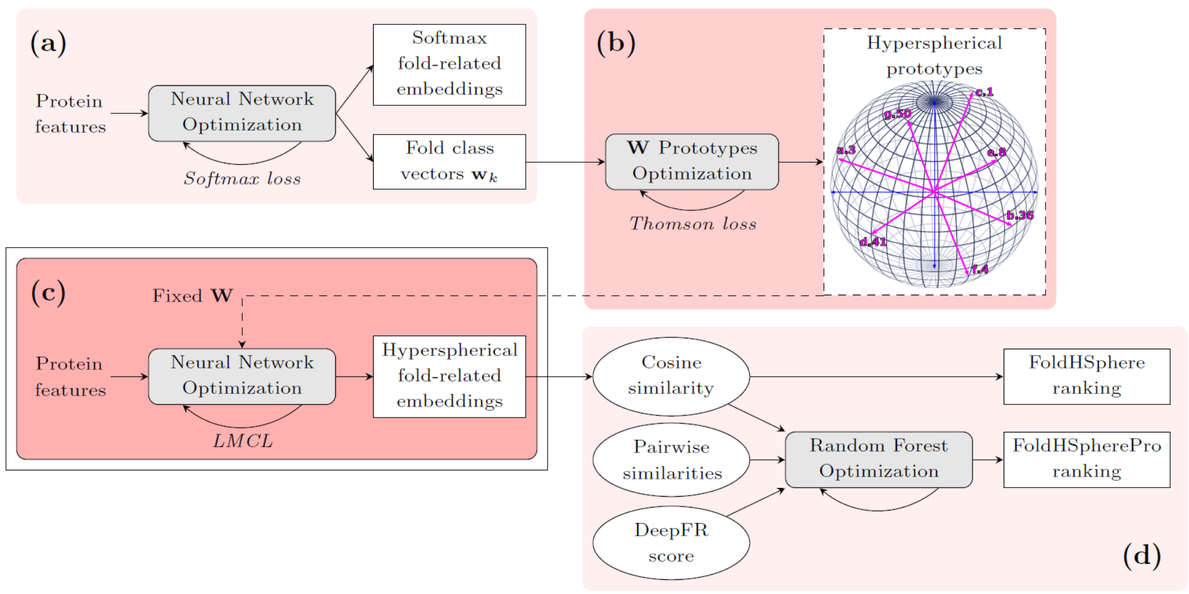
FoldHSphere: Deep Hyperspherical Embeddings for Protein Fold Recognition

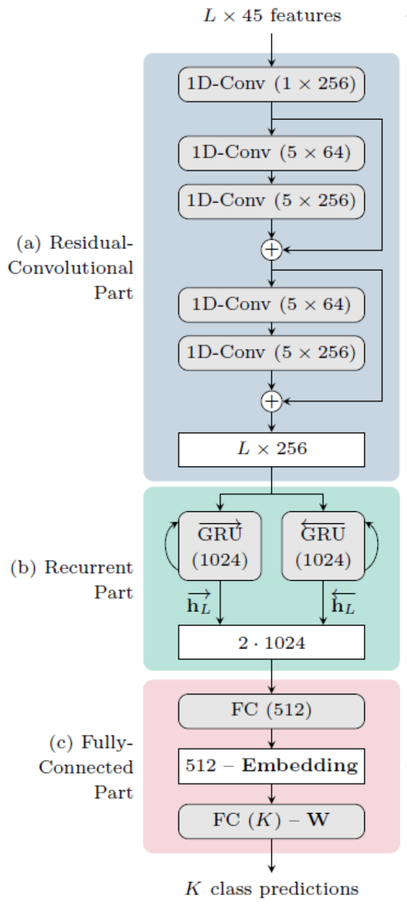
Amelia Villegas-Morcillo, Victoria Sanchez, and Angel M. Gomez

1. INTRODUCTION

Current state-of-the-art deep learning approaches for protein fold recognition learn protein embeddings that improve prediction performance at the fold level. However, there still exists a performance gap at the fold level and the (relatively easier) family level, suggesting that it might be possible to learn an embedding space that better represents the protein folds. In this paper, we propose the FoldHSphere method towards this goal through a two-stage learning procedure. We first obtain prototype vectors for each fold class that are maximally separated in hyperspherical space. We then train a neural network by minimizing the angular large margin cosine loss (LMCL) to learn protein embeddings clustered around the corresponding hyperspherical fold prototypes. Our network architectures, ResCNN-GRU and ResCNN-BGRU, process the input protein sequences by applying several residual-convolutional blocks followed by a gated recurrent unit-based recurrent layer. Evaluation results on the LINDAHL dataset indicate that the use of our hyperspherical embeddings effectively bridges the performance gap at the family and fold levels. Furthermore, our FoldHSpherePro ensemble method outperforms the current state-of-the-art.

2. FOLDHSPHERE METHOD





2a) Softmax Training

- **Train to classify protein domains into K folds**
  - SCOPe 2.06 training dataset: ~16k samples from K =1154 folds
  - Lx45 input features: one-hot encoding of amino acids + PSSM + secondary structure + solvent accessibility
- **ResCNN-BGRU** neural network model
- **Softmax cross-entropy loss:**

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{w}_{y_i}^T\mathbf{x}_i}}{\sum_{k=1}^{K}e^{\mathbf{w}_k^T\mathbf{x}_i}}$$

2b) Prototype Optimization

- Optimize the distribution of the K fold classification vectors $\mathbf{w}_k$:
  - Maximally separate $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ in the hyperspherical space
  - $\mathbf{W}^{softmax}$ contains a suitable initial arrangement of the fold prototypes
- **Thomson Loss (THL)** [1]:

$$L_{th} = \sum_{k=1}^{K}\sum_{j=1}^{k-1}\left\|\frac{\mathbf{w}_k}{\|\mathbf{w}_k\|} - \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}\right\|_2^{-2}$$

2c) LMCL Training

- **Train ResCNN-BGRU model**
  - Use hyperspherical prototypes as a fixed non-trainable classification matrix **W**
  - Extract 512-dim hyperspherical embeddings
- **Large margin cosine loss (LMCL)** [2]:

$$L_{lmc} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i,i}\right)-m\right)} + \sum_{k\neq y_i}e^{s\cos\left(\theta_{k,i}\right)}}$$

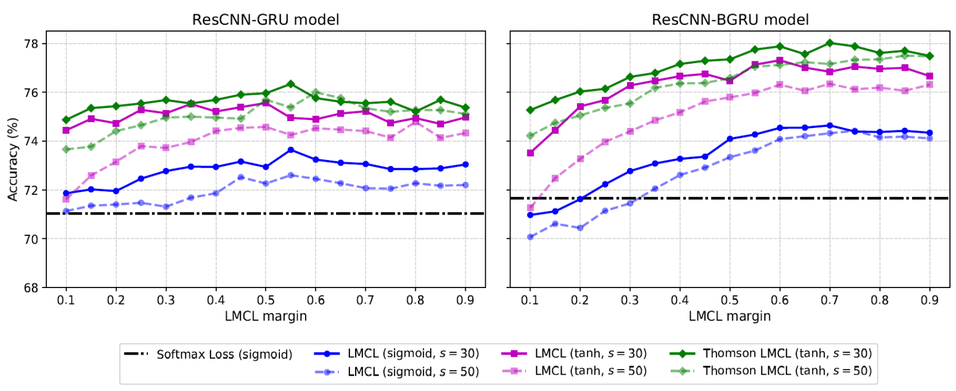
  - s (scale) and m (margin) hyperparameters

2d) Scoring and Fold Recognition

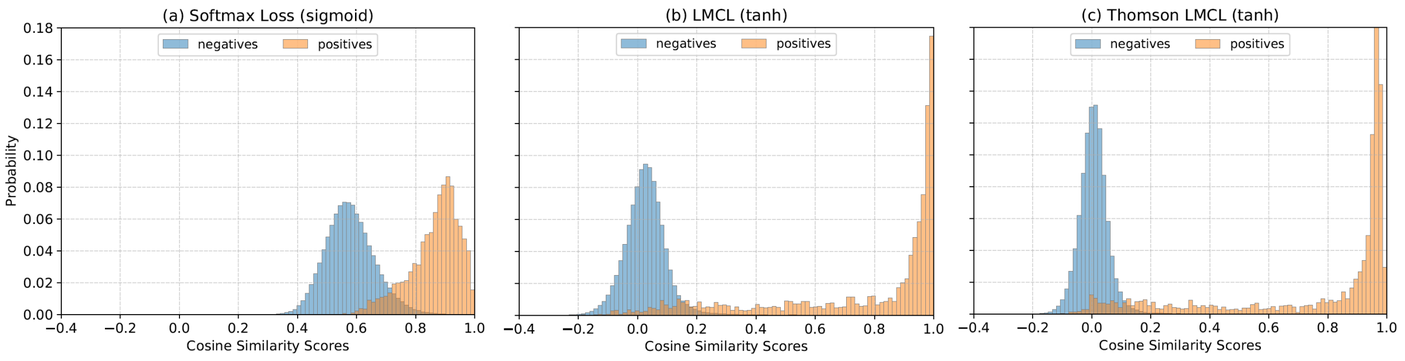
- **LINDAHL test set**
  - 976 protein domains from 320 folds
- **Cosine similarity scores**
  - Between each pair of samples (using hyperspherical embeddings)
- **Top-1 / Top-5 ranking accuracy**
  - Family, Superfamily and Fold levels from SCOP
- **FoldHSpherePro (ensemble)**
  - Random Forest model (samples from the same or different fold)
  - Input: FoldHSphere score + 84 pairwise similarities [3] + DeepFR score [4]
  - 10-stage cross-validation over LINDAHL

3. EXPERIMENTAL RESULTS

- **5-stage cross-validation fold classification results**
  - Models: ResCNN-GRU and ResCNN-BGRU
  - Embedding layer: sigmoid or tanh activation function
  - Softmax Loss, LMCL (end-to-end), Thomson LMCL (hyperspherical prototypes)
  - LMCL: scales $s = \{30, 50\}$ and margins $m = [0.1, \ldots, 0.9]$



- **LINDAHL cosine similarity histograms**



- **Selected model:**
  - ResCNN-BGRU trained with Thomson LMCL ($s = 30$, $m = 0.6$)
- **LINDAHL fold recognition results**

Method	Family		Superfamily		Fold	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
RF-Fold [3]	84.5	91.5	63.4	79.3	40.8	58.3
DN-Fold [3]	84.5	91.2	61.5	76.5	33.6	60.7
RFDN-Fold [3]	84.7	91.5	65.7	78.8	37.7	61.7
DeepFR [4]	65.4	83.4	51.4	67.1	56.1	70.1
CNN-BGRU [5]	71.0	87.7	60.1	77.2	58.3	78.8
FoldHSphere	76.4	89.2	72.8	86.4	75.1	84.1
DeepFRpro [4]	83.1	92.3	69.6	82.5	66.0	78.8
CNN-BGRU-RF+ [5]	**85.4**	**93.5**	73.3	87.8	76.3	85.7
FoldHSpherePro	85.2	93.0	**79.0**	**89.2**	**81.3**	**90.3**

4. CONCLUSIONS

- ❑ The proposed methodology allows us to **learn a better fold embedding space** and thus extract discriminative embeddings for the protein domains
- ❑ **FoldHSphere** alone provides a remarkable performance boost at the superfamily and fold levels
- ❑ Our **FoldHSpherePro** ensemble method significantly improves the state-of-the-art results
- ❑ The **hyperspherical embeddings** are effective at finding template proteins, even when the amino acid sequence similarities are low

References

[1] Thomson (1904). XXIV. On the structure of the atom. *Philosophical Magazine*
[2] Wang et al. (2018). CosFace: Large margin cosine loss for deep face recognition. *CVPR Proceedings*
[3] Jo et al. (2015). Improving protein fold recognition by deep learning networks. *Scientific Reports*
[4] Zhu et al. (2017). Improving protein fold recognition by extracting fold-specific features from predicted residue-residue contacts. *Bioinformatics*
[5] Villegas-Morcillo et al. (2020). Protein fold recognition from sequences using convolutional and recurrent neural networks. *IEEE/ACM TCBB*

Funding

- Spanish Ministry of Science, Innovation and Universities Project No. PID2019-104206GB-I00 / SRA (State Research Agency) / 10.13039/501100011033
- FPI grant BES-2017-079792

Contact: ameliavm@ugr.es