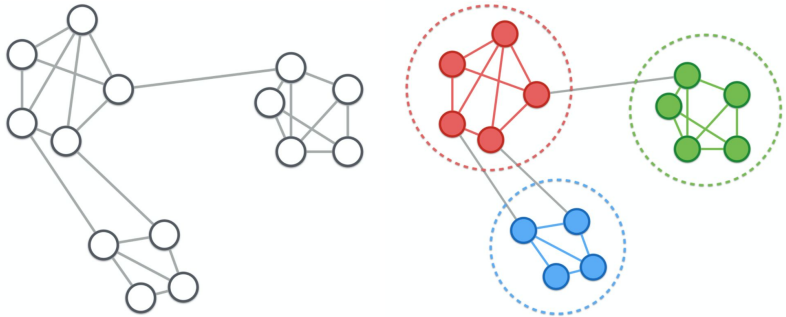
seqSCAN: Unsupervised Classification of Proteins for New Function Discovery

Meet Barot, Vladimir Gligorijevic, Kyunghyun Cho, Richard Bonneau

Unsupervised learning is crucial for biological data

- Most data in biology is unlabeled
  - Experiments are expensive, error-prone
  - Supervised learning algorithms are limited by this
- Most label sets are incomplete
- We need tools to get functional categories of proteins whether they have labels or not
- Discovery of these categories can enable us to infer new GO terms, correct old terms, and create entire ontologies



https://blog.insightdatascience.com/graph-based-machine-learning-6e2bd8926a0

Clustering with neural networks

- A neural network trained to cluster proteins would be able to be trained batchwise and have constant time inference for a new protein
  - For methods like spectral clustering, inference requires computing pairwise similarities to previous samples
- Using class activation maps, we would be able to highlight specific parts of the protein that correspond to these hypothetical classes
- Feature learning can be combined with the clustering process in neural networks

Learning to classify images without labels (SCAN)

Step 1: Self-supervised model to learn feature space

$$\min_{\theta} d(\Phi_{\theta}(X_i), \Phi_{\theta}(T[X_i])).$$

Van Gansbeke, Wouter, et al. "Learning To Classify Images Without Labels." *arXiv preprint arXiv:2005.12320* (2020).

Learning to classify images without labels (SCAN)

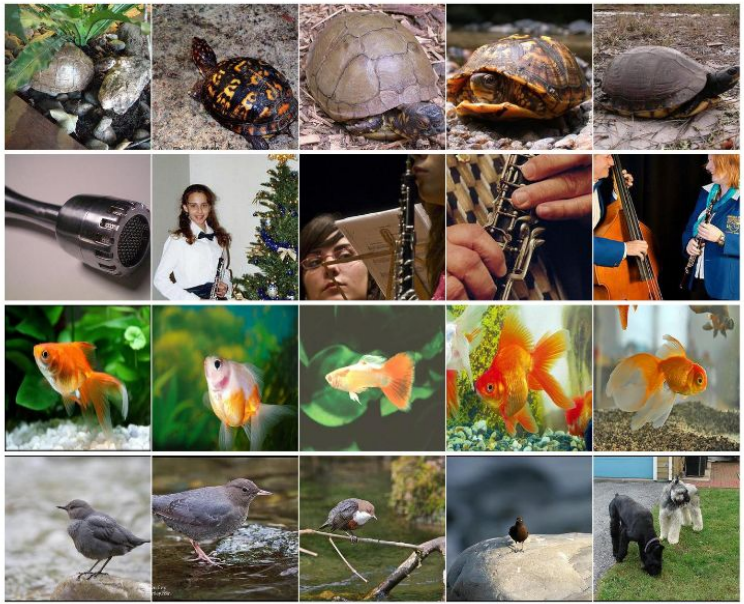


Fig. 1: Images (first column) and their nearest neighbors (other columns) [51].

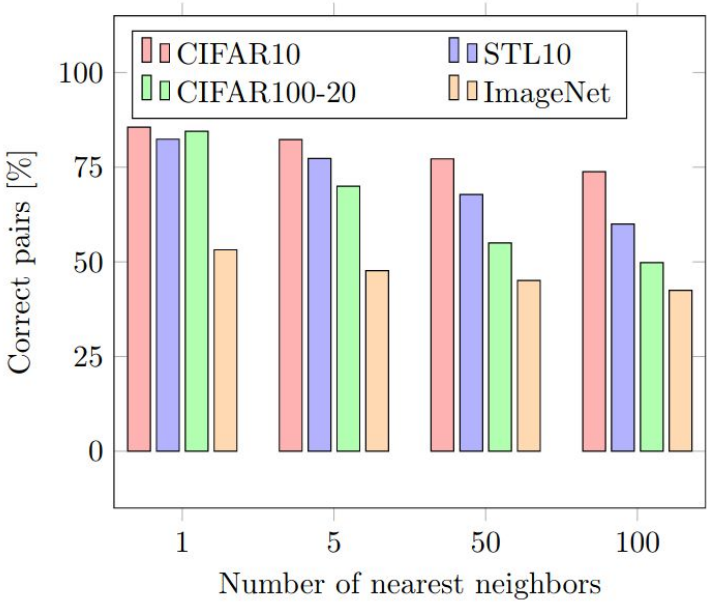


Fig. 2: Neighboring samples tend to be instances of the same semantic class.

Learning to classify images without labels (SCAN)

Step 2: Train previous neural network, now with a softmax output, with the following loss:

$$\Lambda = -\frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \sum_{k \in \mathcal{N}_X} \log \langle \Phi_\eta(X), \Phi_\eta(k) \rangle + \lambda \sum_{c \in \mathcal{C}} \Phi'^c_\eta \log \Phi'^c_\eta,$$

$$\text{with } \Phi'^c_\eta = \frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \Phi^c_\eta(X).$$

Van Gansbeke, Wouter, et al. "Learning To Classify Images Without Labels." *arXiv preprint arXiv:2005.12320* (2020).

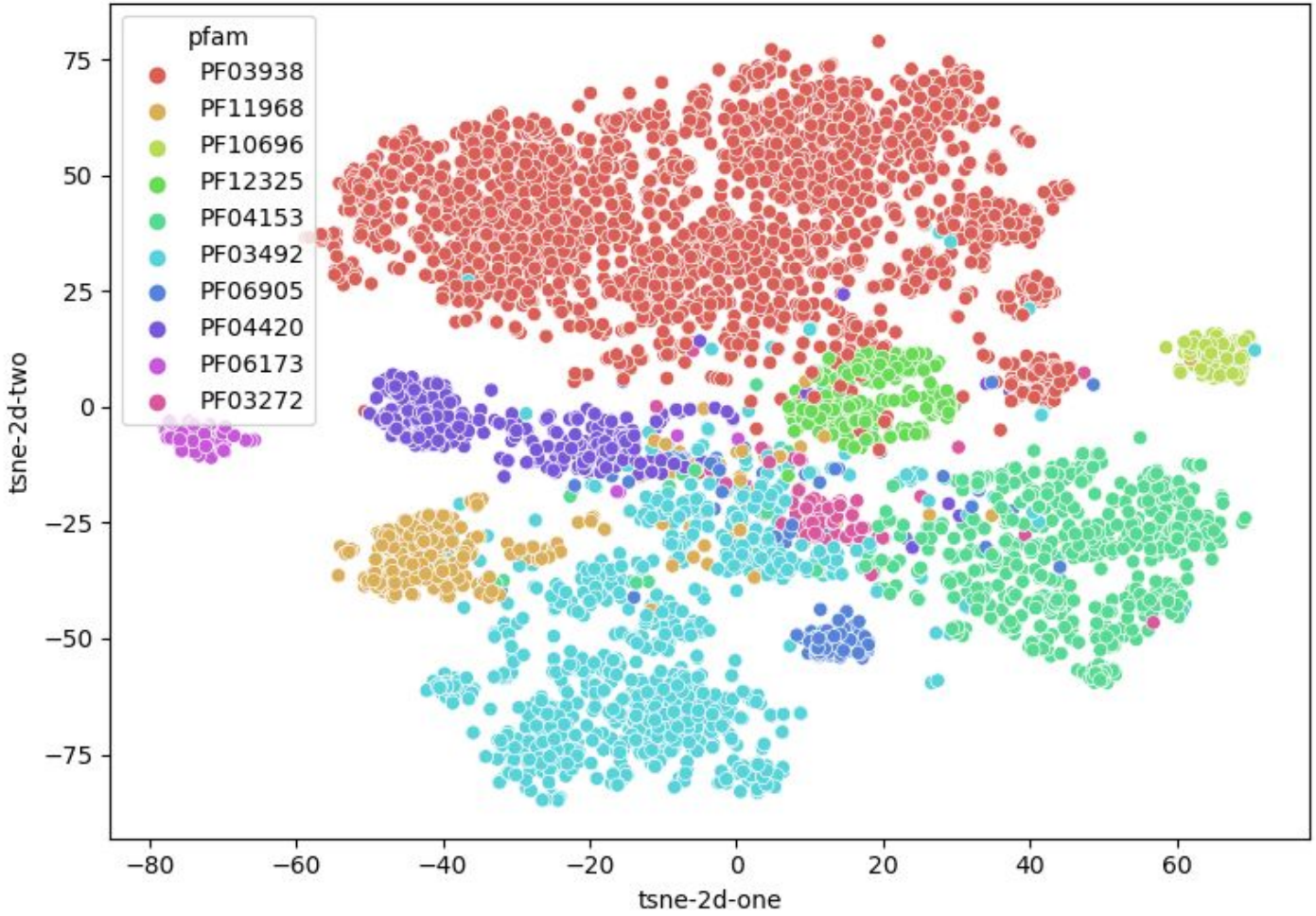
Learning to classify images without labels (SCAN)

Dataset	**CIFAR10**			**CIFAR100-20**			**STL10**		
Metric	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Supervised	93.8	86.2	87.0	80.0	68.0	63.2	80.6	65.9	63.1
Pretext [7] + K-means	65.9 ± 5.7	59.8 ± 2.0	50.9 ± 3.7	39.5 ± 1.9	40.2 ± 1.1	23.9 ± 1.1	65.8 ± 5.1	60.4 ± 2.5	50.6 ± 4.1
SCAN* (Avg $\pm$ Std)	81.8 ± 0.3	71.2 ± 0.4	66.5 ± 0.4	42.2 ± 3.0	44.1 ± 1.0	26.7 ± 1.3	75.5 ± 2.0	65.4 ± 1.2	59.0 ± 1.6
SCAN† (Avg $\pm$ Std)	87.6 ± 0.4	78.7 ± 0.5	75.8 ± 0.7	45.9 ± 2.7	46.8 ± 1.3	30.1 ± 2.1	76.7 ± 1.9	68.0 ± 1.2	61.6 ± 1.8
SCAN† (Best)	**88.3**	**79.7**	**77.2**	**50.7**	**48.6**	**33.3**	**80.9**	**69.8**	**64.6**
SCAN† (Overcluster)	86.2 ± 0.8	77.1 ± 0.1	73.8 ± 1.4	55.1 ± 1.6	50.0 ± 1.1	35.7 ± 1.7	76.8 ± 1.1	65.6 ± 0.8	58.6 ± 1.6

Pfam Experiment --- seqSCAN

- Using self-supervised sequence model to extract useful features from sequence (language model trained on 10 million protein sequences from Pfam)
- Dataset of 10 protein families, total ~6000 proteins
- Evaluate clusters obtained using Normalized Mutual Information (NMI) with respect to protein family labels

$$\mathbf{NMI}(\boldsymbol{\Omega}, \mathbf{C}) = \frac{I(\boldsymbol{\Omega}; \mathbf{C})}{[H(\boldsymbol{\Omega}) + H(\mathbf{C})]/2}$$


pfam
PF03938
PF11968
PF10696
PF12325
PF04153
PF03492
PF06905
PF04420
PF06173
PF03272
75
50
25
0
−25
−50
−75
tsne-2d-two
−80
−60
−40
−20
0
20
40
60
tsne-2d-one

Using SCAN loss to cluster proteins using learned features

- Train a single-layer model on the learned features with softmax output with the SCAN loss function:

$$\Lambda = -\frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \sum_{k \in \mathcal{N}_X} \log \langle \Phi_\eta(X), \Phi_\eta(k) \rangle + \lambda \sum_{c \in \mathcal{C}} \Phi'^c_\eta \log \Phi'^c_\eta,$$

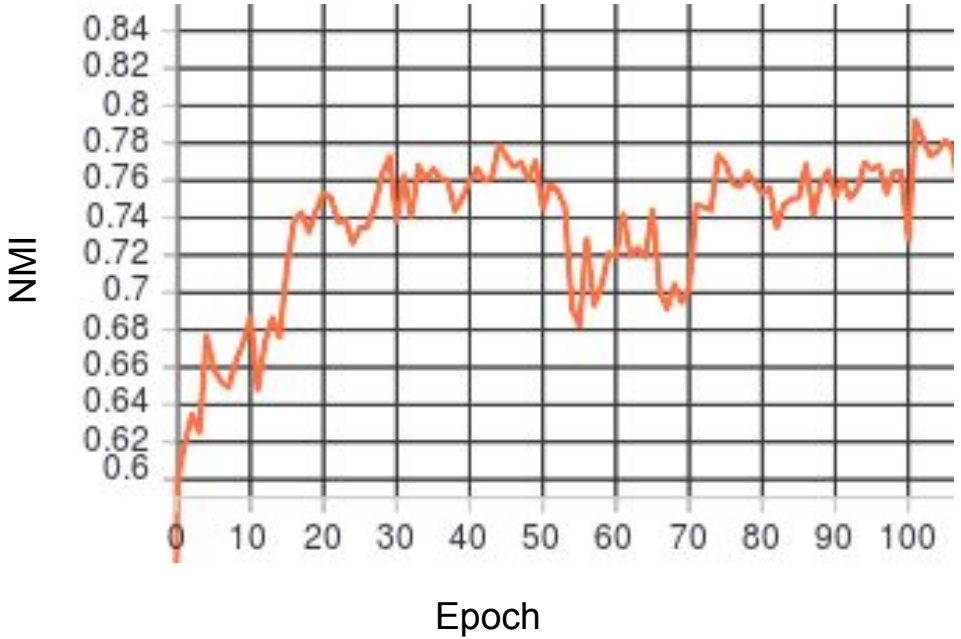
$$\text{with } \Phi'^c_\eta = \frac{1}{|\mathcal{D}|} \sum_{X \in \mathcal{D}} \Phi^c_\eta(X).$$

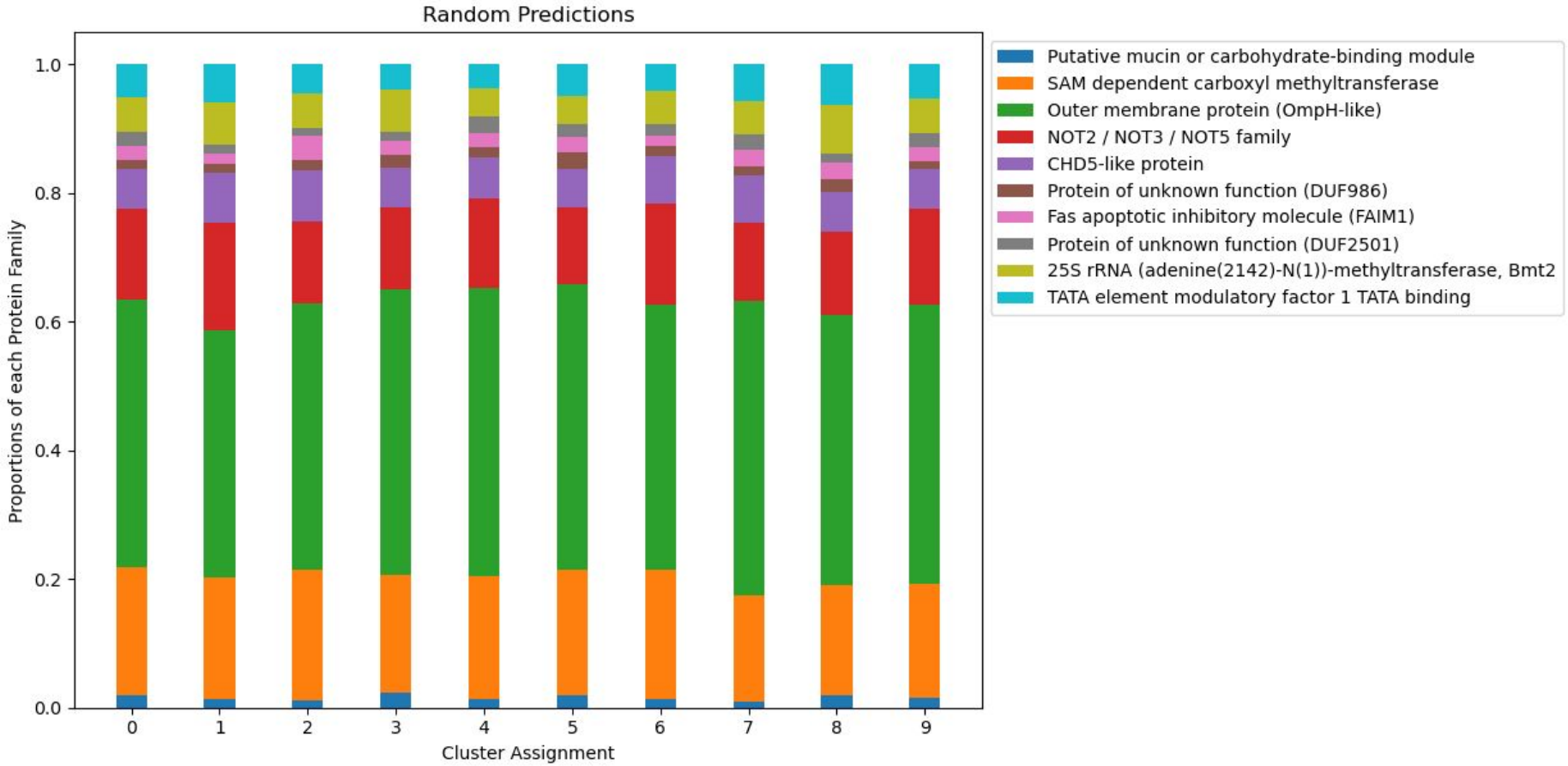
NMI during “SeqSCAN” Training

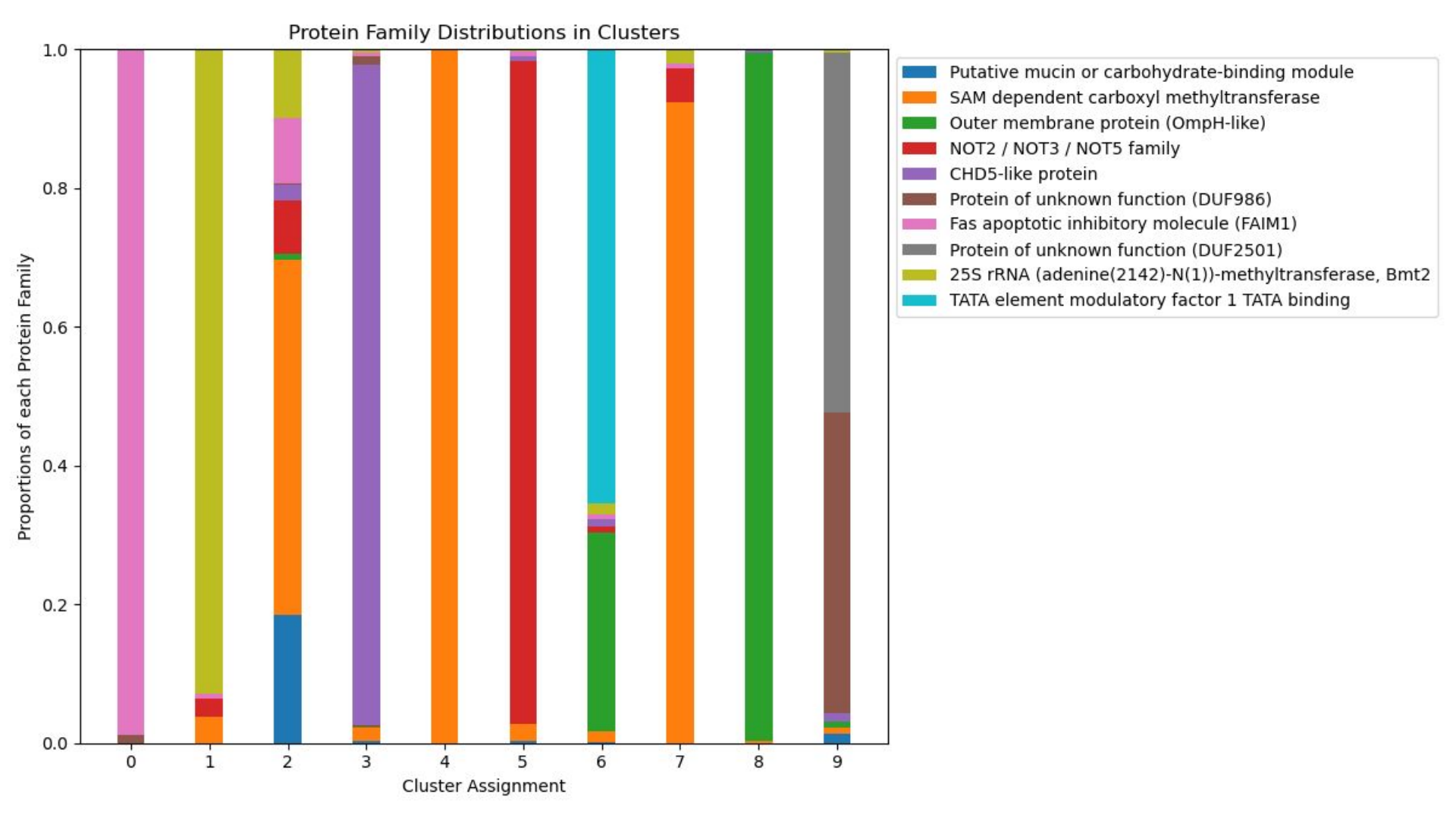
Baseline:
PCA+K-means on sequence features

- NMI: 0.61

$$\mathrm{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2}$$

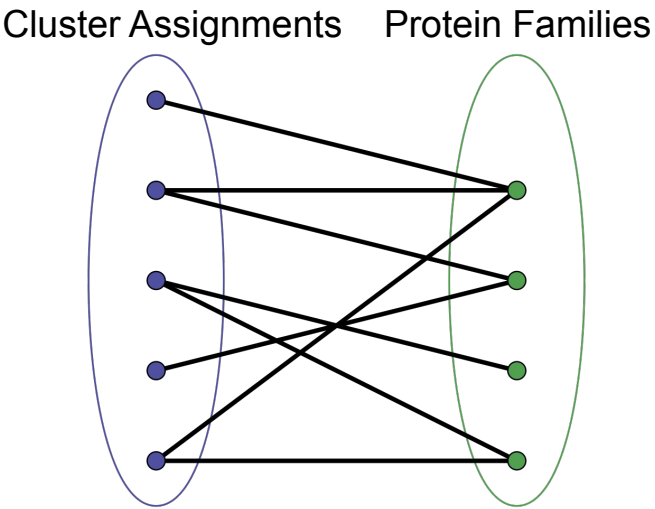


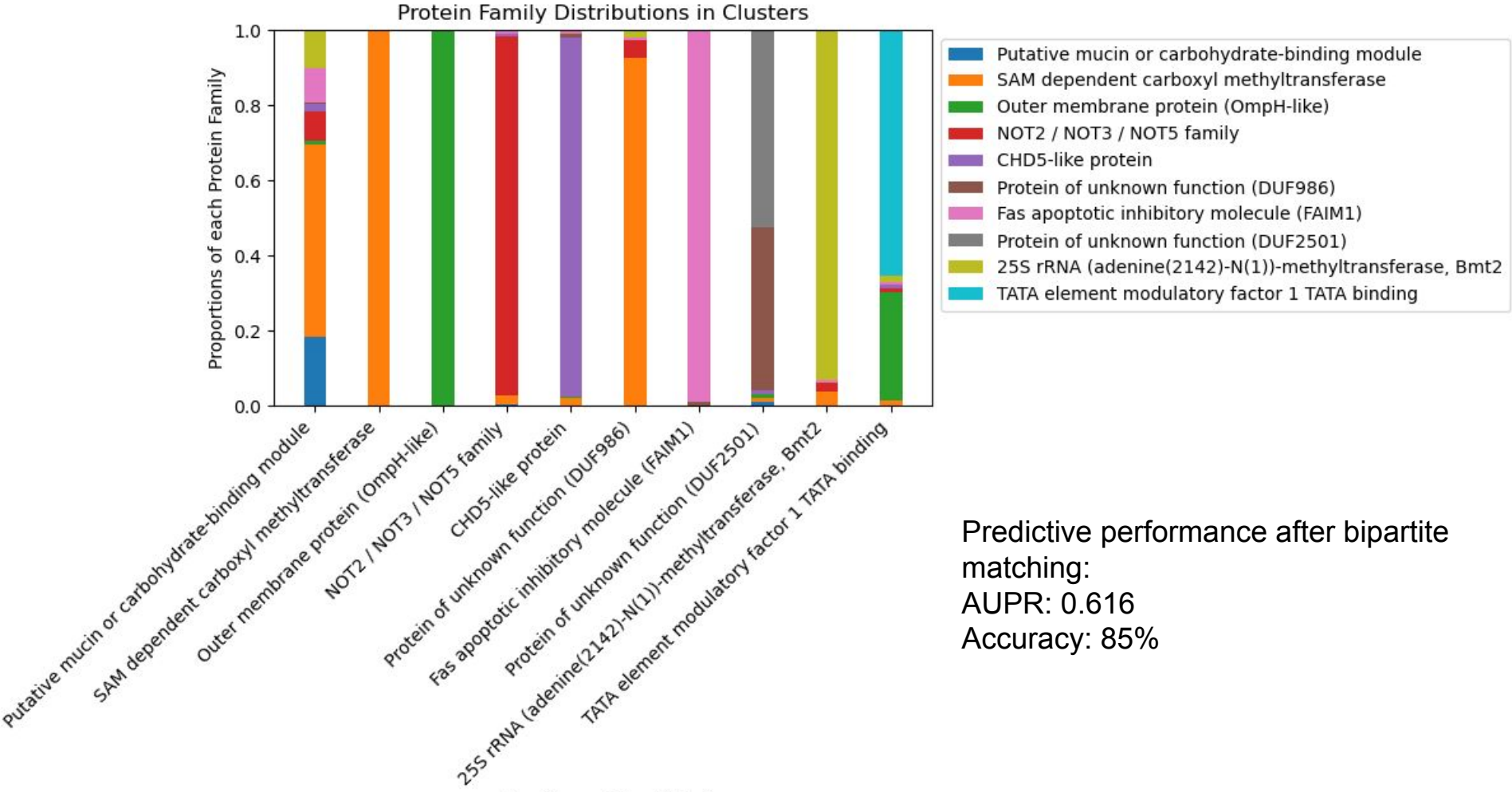

Random Predictions
Proportions of each Protein Family
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
Cluster Assignment
Putative mucin or carbohydrate-binding module
SAM dependent carboxyl methyltransferase
Outer membrane protein (OmpH-like)
NOT2 / NOT3 / NOT5 family
CHD5-like protein
Protein of unknown function (DUF986)
Fas apoptotic inhibitory molecule (FAIM1)
Protein of unknown function (DUF2501)
25S rRNA (adenine(2142)-N(1))-methyltransferase, Bmt2
TATA element modulatory factor 1 TATA binding


Protein Family Distributions in Clusters
Proportions of each Protein Family
Cluster Assignment
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8
9
Putative mucin or carbohydrate-binding module
SAM dependent carboxyl methyltransferase
Outer membrane protein (OmpH-like)
NOT2 / NOT3 / NOT5 family
CHD5-like protein
Protein of unknown function (DUF986)
Fas apoptotic inhibitory molecule (FAIM1)
Protein of unknown function (DUF2501)
25S rRNA (adenine(2142)-N(1))-methyltransferase, Bmt2
TATA element modulatory factor 1 TATA binding

Matching cluster assignments with labels

- Bipartite matching: maximizing accuracy of the cluster assignments with respect to their protein families





Predictive performance after bipartite matching:
AUPR: 0.616
Accuracy: 85%

Scaling up to PfamA

- Training clustering model on 16 million proteins (15k Pfam families)
- Test on 1.8 million proteins (13k families)

	PCA+K-means	seqSCAN
Training NMI	0.499	**0.516**
Test NMI	0.523	**0.541**

Vladimir Gligorijević
Research Scientist

Richard Bonneau

Kyunghyun Cho

The Bonneau lab!

www.meetbarot.com
meetbarot@nyu.edu