Distilling ESM-1b: A Compact Protein Language Model

Sergio G. Charles

Stats 326 Spring 2025

Abstract

We distill the 669M-parameter ESM-1b protein language model into smaller Transformers (1M–33M params). Student models are trained using a combination of cross-entropy and distillation loss to align with the teacher's soft predictions. While perplexity gains are modest, distillation improves downstream Pfam and SCOPe classification considerably.

Key Result

A 33.7M parameter student trained with distillation at T=2.0 **outperforms** ESM-1b on SCOPe structural classification (F1 = 0.71 vs. 0.67), despite being 20× smaller.

Distillation significantly improves biological annotation (Pfam, SCOPe) classification metrics compared to student models only trained using cross-entropy, even when perplexity gains are minimal.

Methods

We distill the 669M parameter ESM-1b protein language model into a family of smaller Transformer students using masked language modeling (MLM) and knowledge distillation (KD).

Model Architecture

- **Teacher:** 33 layers, 1280 hidden dim, 20 attention heads.
- **Students:**
  - Small (1.2M): 8 layers, 128 dim, 2 heads
  - Medium (5.6M): 10 layers, 256 dim, 4 heads
  - Large (33.7M): 28 layers, 384 dim, 6 heads

All models share a vocabulary and are trained on 50k protein sequences (UniRef50), with Pfam and SCOPe labels.

Training Objective

Each protein sequence $x = (x_1, \ldots, x_T)$ is corrupted into $\tilde{x}$ by masking 15% of tokens. The standard MLM loss is:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|}\sum_{i \in M} \log p_\theta(x_i|\tilde{x})$$

where $p_\theta(x_i|\tilde{x}) = \text{softmax}(z_i)$

Distillation

We match student and teacher softmax outputs at temperature T:

$$\mathcal{L}_{\text{KD}} = \frac{1}{|M|}\sum_{i \in M} \mathrm{D}_{\text{KL}}\left(\text{softmax}\left(\frac{z_i^t}{T}\right) \,||\, \text{softmax}\left(\frac{z_i^s}{T}\right)\right)$$

The total loss is a weighted sum:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{MLM}} + \lambda T^2 \mathcal{L}_{\text{KD}}$$

Optimization

- Optimizer: AdamW, learning rate 10e-3
- Batch sizes: 512 (small), 128 (medium), 64 (large)
- Epochs: 10 (small), 20 (medium), 30 (large)
- Temperatures: T∈{0.5, 1.0, 2.0}
- Used $\lambda = 0.5$ to balance CE/distillation.

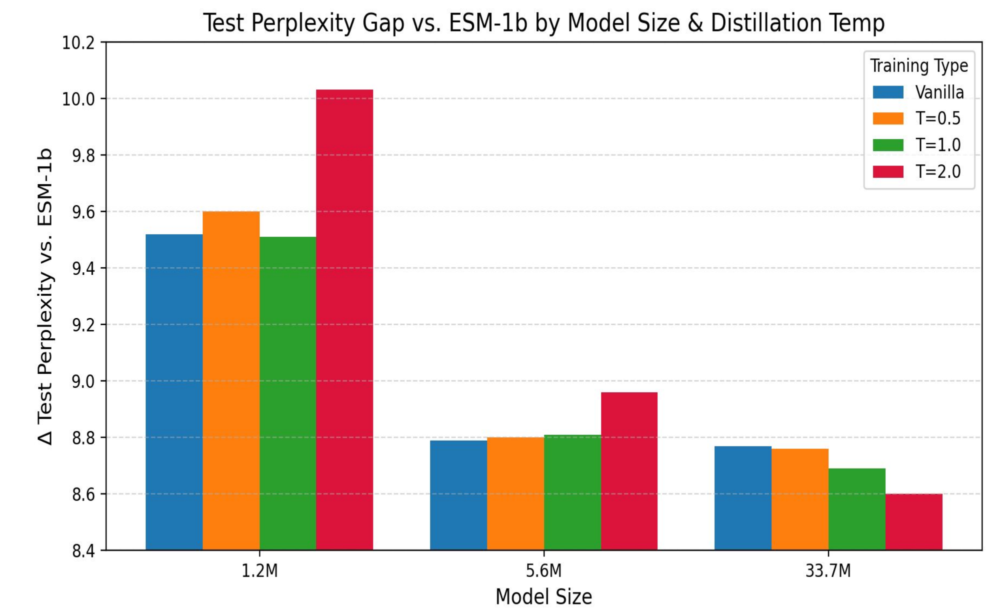
Teacher weights are frozen. Models were trained on NVIDIA GH200 Superchip over 36 hours.

Distillation Results

Model	Params (M)	Compression Rate (×)	Train PPL	Test PPL	ΔTest PPL
ESM-1b	669.2	1.0	4.64	4.70	0.00
2-gram	0.0005	1.3M	18.08	18.08	13.38
3-gram	0.01	69708.3	17.91	17.95	13.25
4-gram	0.17	3891.9	17.53	**17.76**	**13.06**
8-gram	15.2	44.01	11.46	20.90	16.20
Student-h128-L8-VANILLA	1.2	557.7	14.18	14.22	9.52
Student-h128-L8-T0.5	1.2	557.7	14.26	14.30	9.60
Student-h128-L8-T1.0	1.2	557.7	14.22	**14.21**	**9.51**
Student-h128-L8-T2.0	1.2	557.7	14.73	14.73	10.03
Student-h256-L10-VANILLA	5.6	119.5	13.44	**13.49**	**8.79**
Student-h256-L10-T0.5	5.6	119.5	13.47	13.50	8.80
Student-h256-L10-T1.0	5.6	119.5	13.50	13.51	8.81
Student-h256-L10-T2.0	5.6	119.5	13.74	13.66	8.96
Student-h384-L28-VANILLA	33.7	19.9	13.41	13.47	8.77
Student-h384-L28-T0.5	33.7	19.9	13.37	13.46	8.76
Student-h384-L28-T1.0	33.7	19.9	13.36	13.39	8.69
Student-h384-L28-T2.0	33.7	19.9	13.16	**13.30**	**8.60**

Takeaways

- Student models do much better than n-gram baselines: the best 1.2M model reaches 14.21, while the 33.7M model trained with temperature T=2.0 achieves 13.30, just +8.60 over ESM-1b despite a 19.3x compression.
- Distillation improves performance slightly for 1.2M and 33.7M models, but doesn't help for 5.6M.



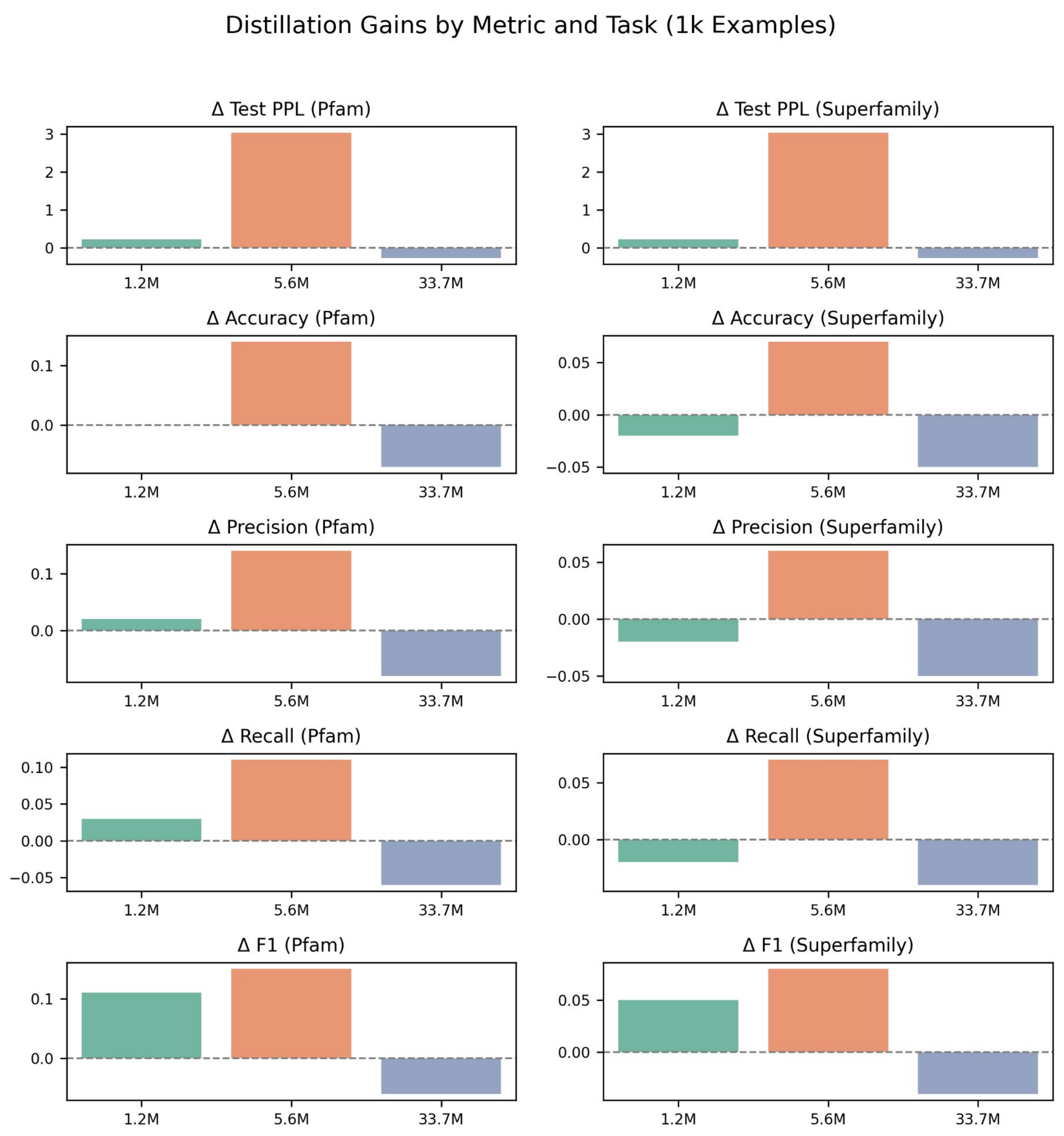
Hypothesis

$$\Delta_{\text{Distill}}(M, D) \propto \frac{1}{M \cdot D}$$

where M is model size and D is dataset size.

Low-data regime with 1k-example training set:

- Distillation is most effective for mid-sized students in low-data regimes, improving both perplexity and downstream performance.



Embedding Analysis

Model	Accuracy	Precision	Recall	F1
ESM-1b	**0.80**	**0.81**	**0.80**	**0.80**
Student-h128-L8-VANILLA	0.75	0.75	0.75	0.73
Student-h128-L8-T0.5	0.76	0.75	0.76	0.74
Student-h128-L8-T1.0	0.72	0.70	0.72	0.70
Student-h128-L8-T2.0	0.32	0.31	0.32	0.31
Student-h256-L10-VANILLA	0.73	0.71	0.73	0.71
Student-h256-L10-T0.5	0.72	0.73	0.72	0.71
Student-h256-L10-T1.0	0.69	0.68	0.69	0.66
Student-h256-L10-T2.0	0.66	0.64	0.66	0.31
Student-h384-L28-VANILLA	0.70	0.71	0.70	0.69
Student-h384-L28-T0.5	0.72	0.72	0.72	0.70
Student-h384-L28-T1.0	0.70	0.68	0.70	0.68
Student-h384-L28-T2.0	0.79	0.80	0.79	0.79

Table: Pfam classification performance.

Model	Accuracy	Precision	Recall	F1
ESM-1b	0.68	0.70	0.68	0.67
Student-h128-L8-VANILLA	0.58	0.59	0.59	0.58
Student-h128-L8-T0.5	0.55	0.57	0.55	0.55
Student-h128-L8-T1.0	0.54	0.54	0.54	0.53
Student-h128-L8-T2.0	0.32	0.32	0.32	0.30
Student-h256-L10-VANILLA	0.59	0.57	0.59	0.58
Student-h256-L10-T0.5	0.54	0.57	0.58	0.53
Student-h256-L10-T1.0	0.53	0.54	0.53	0.53
Student-h256-L10-T2.0	0.54	0.54	0.54	0.53
Student-h384-L28-VANILLA	0.58	0.57	0.58	0.57
Student-h384-L28-T0.5	0.58	0.58	0.58	0.57
Student-h384-L28-T1.0	0.59	0.58	0.59	0.58
Student-h384-L28-T2.0	**0.72**	**0.71**	**0.72**	**0.71**

Table: SCOPe Superfamily classification performance.

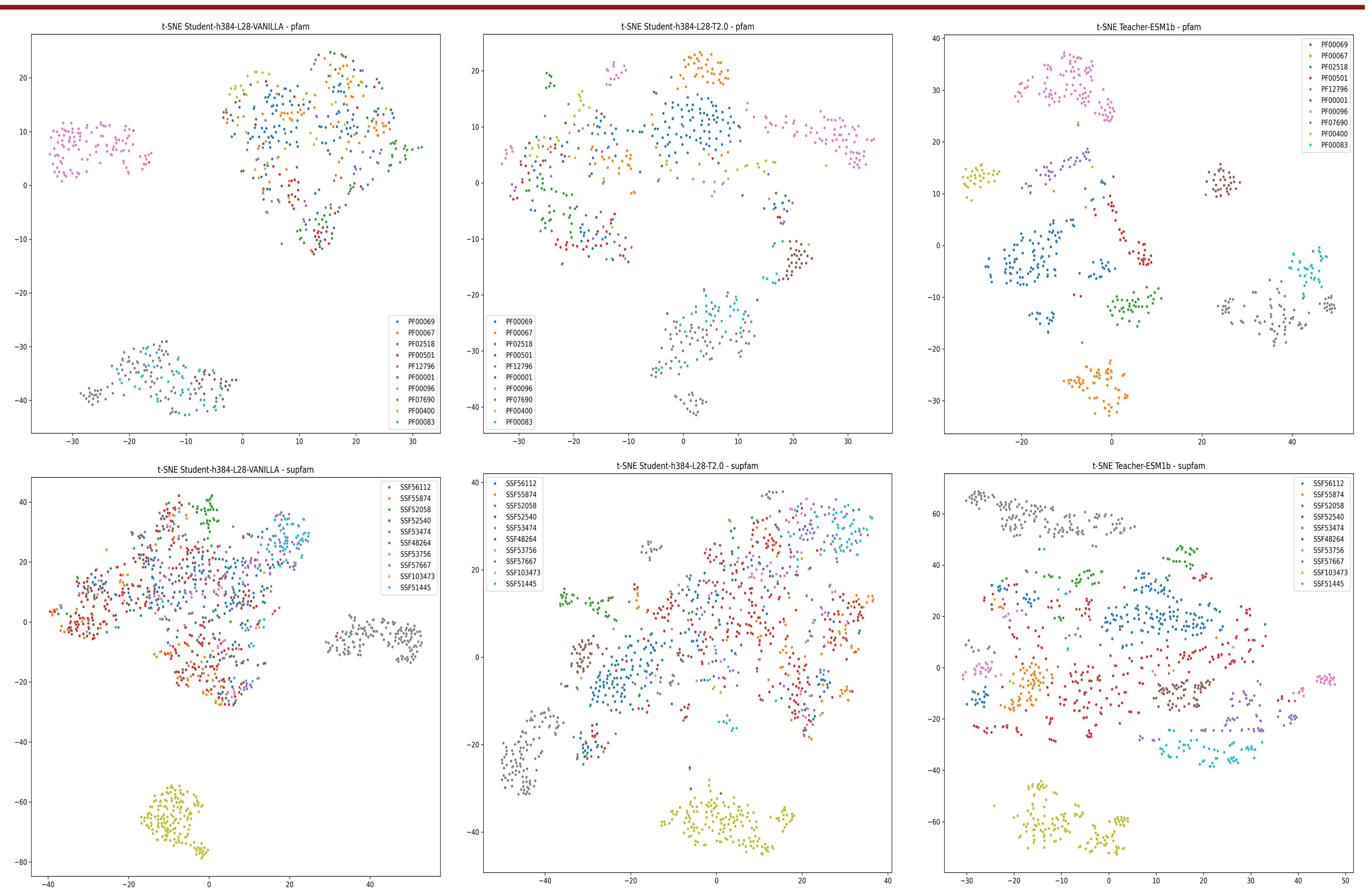


Figure: t-SNE plots of CLS token embeddings for Pfam (top row) and Superfamily (bottom row) labels. Each column corresponds to a different model: Student-h384-L8-VANILLA (left), Student-h384-L8-T=2.0 (middle), and ESM-1b (right).