Distilling ESM-1b: A Compact Protein Language Model via Knowledge Distillation

Sergio Charles

Abstract

We distill the ESM-1b protein language model into a family of smaller Transformer models. ESM-1b is a high-capacity 669M parameter model that captures rich biological information from sequences [1]. We construct nine student models of varying sizes, from 1M to 33M parameters, and train them with and without knowledge distillation. The distilled models are optimized to mimic the teacher's output distribution, in addition to learning from true labels. We evaluate all models on Pfam and structural superfamily SCOPe classification tasks using k-nearest-neighbor classifiers in embedding space, reporting accuracy, precision, recall, and F1 for each model. Our results show that larger models achieve lower perplexity and better classification metrics. While distillation does not have a significant effect on reducing perplexity, we find it noticeably improves protein annotation classification over vanilla training. We also find distillation to be useful for low-capacity student models in data-scarce regimes.

1 Introduction

Protein language models (PLMs) have emerged as powerful tools for extracting evolutionary and structural information from sequences [1]. Meta AI's ESM-1b model is a 669M-parameter Transformer trained on 250M protein sequences from UniParc [2], and its learned representations encode biochemical properties, secondary/tertiary structure, and evolutionary homology [1]. These contextual embeddings cluster residues by chemical property and capture family relationships, leading to state-of-the-art predictions of mutational effects and contacts [1]. However, such large models are expensive to deploy.

Knowledge distillation provides a strategy to compress a high-capacity model into a smaller, student model by training it to mimic the teacher's logits [3] [4]. In NLP, distilled models like DistilBERT (40% smaller) retain $\sim$97% of BERT's capabilities while being 40% smaller and 60% faster [4]. Recent work has begun applying distillation to PLMs (e.g., ProtGPT2 [5], ProtBERT [6] variants), but systematic studies of Transformer distillation for proteins are limited. We aim to fill this gap by distilling ESM-1b into student Transformers of three sizes. ESM-1b was pre-trained on UniRef50, a clustering of UniParc at 50% sequence identity [1] dataset with approximately 30 million proteins. We train on a 50k subset of the UniRef50 using both the standard cross-entropy loss and a soft-target distillation loss [3].

2 Methods

The teacher model is ESM-1b (669M parameters) with 34 Transformer encoder layers, 1280 hidden dimension, 5120 intermediate MLP dimension, and 20 attention heads [1]. We derive 12 student models of a smaller size. Table 1 lists the number of parameters for each model. We broadly categorize them by size:

- Teacher (669M parameters): 33 layers, 1280 hidden dimension, 5120 intermediate dimension, 20 attention heads.
- Large student (33.7M parameters): 28 layers, 384 hidden dimension, 256 intermediate dimension, and 6 attention heads.
- Medium student (5.6M parameters): 10 layers, 256 hidden dimension, 512 intermediate dimension, and 4 attention heads.
- Small student (1.2M parameters): 8 layers, 128 hidden dimension, 768 intermediate dimension, and 2 attention heads.

All models use the same vocabulary and masked language modeling (MLM) objective as ESM-1b. We curated a dataset of 50k UniRef50 protein sequences with Pfam and superfamily SCOPe annotations, and used a 80-10-10 train-val-test split. We compare two training regimes: (1) vanilla training with true-token cross-entropy loss, and (2) distillation training. For distillation, we use the ESM-1b teacher to produce soft logits on each masked token. We match the student's softmax output (at temperature T) to the teacher's soft targets via a KL divergence term, in addition to the cross-entropy loss. This transfers the teacher's "dark knowledge", i.e. relative probabilities of wrong tokens, to the student [3]. Student models use the same BERT-style Transformer architecture as ESM-1b. Training was done on a Lambda cloud GH200 Superchip for $10, 20,$ and 30 epochs for small, medium, and large models, respectively. We did a sweep over $T = 0.5, 1.0, 2.0$ for all models. Running all 17 model variants took $\sim$36 hours.

Preprint. Under review.

Masked Language Modeling Objective. The ESM-1b model is trained using the *masked language modeling* (MLM) objective, adapted from BERT [6] and applied to protein sequences. Given a protein sequence $x = (x_1, x_2, \ldots, x_T)$, where each x_t is an amino acid token from a vocabulary of size $|\mathcal{V}|$, we randomly select a subset $\mathcal{M} \subset \{1, \ldots, T\}$ corresponding to 15% of positions to serve as prediction targets.

For each $i \in \mathcal{M}$, the input token x_i is corrupted as follows:

- 80% of the time, x_i is replaced with a special `[MASK]` token,
- 10% of the time, x_i is replaced with a random token from $\mathcal{V} \setminus \{x_i\}$,
- 10% of the time, x_i is left unchanged.

Let $\tilde{x}$ denote the corrupted sequence. The model processes $\tilde{x}$ and outputs a distribution over the vocabulary at each position. For each masked position $i \in \mathcal{M}$, the model produces logits $z_i \in \mathbb{R}^{|\mathcal{V}|}$, which are transformed via a softmax:

$$p_\theta(x_i \mid \tilde{x}) = \text{softmax}(z_i) \in \mathbb{R}^{|\mathcal{V}|}.$$

The MLM loss is the average cross-entropy between the predicted and true tokens over all masked positions:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{x \sim \mathcal{X}} \, \mathbb{E}_{\mathcal{M}} \left[\sum_{i \in \mathcal{M}} \log p_\theta(x_i \mid \tilde{x}) \right].$$

Each prediction is made over the full vocabulary of size $|\mathcal{V}| = 33$ (including the 20 standard amino acids and special tokens). We evaluate model uncertainty using exponentiated cross-entropy or *perplexity*, defined as:

$$\text{PPL} = \exp\left(\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} - \log p_\theta(x_i \mid \tilde{x}) \right).$$

A PPL of 1 corresponds to perfect prediction, while an PPL near $|\mathcal{V}|$ reflects uniform uncertainty. This formulation encourages the model to learn contextual dependencies between amino acids and capture structural signals from sequence data alone.

3 Distillation

Objective. For each $i \in \mathcal{M}$, let $z_s^{(i)}, z_t^{(i)} \in \mathbb{R}^{|\mathcal{V}|}$ denote the logits from the student and teacher models, respectively, and let $y^{(i)} \in \{0, 1\}^{|\mathcal{V}|}$ be the one-hot vector corresponding to the true amino acid token at position i. The student softmax probabilities at each position are given by:

$$p_s^{(i)} = \text{softmax}(z_s^{(i)}), \quad p_t^{(i)} = \text{softmax}(z_t^{(i)}).$$

The masked cross-entropy loss over the true tokens is:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{c=1}^{|\mathcal{V}|} y_c^{(i)} \log p_{s,c}^{(i)}.$$

For distillation, we soften both distributions using a temperature parameter T:

$$p_s^{(i,T)} = \text{softmax}(z_s^{(i)}/T), \quad p_t^{(i,T)} = \text{softmax}(z_t^{(i)}/T).$$

The distillation loss is the average KL divergence between the teacher and student distributions at the masked positions:

$$\mathcal{L}_{\text{KD}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} D_{\text{KL}}\big(p_t^{(i,T)} \,\|\, p_s^{(i,T)}\big) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \sum_{c=1}^{|\mathcal{V}|} p_{t,c}^{(i,T)} \log \frac{p_{t,c}^{(i,T)}}{p_{s,c}^{(i,T)}}.$$

The total loss is a weighted sum of the two components:

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

where $\lambda \in [0, 1]$ balances the contribution of the distillation signal. Following [3], we include the T^2 scaling factor to match the magnitude of gradients.

Training setup. In all distillation experiments, we set $\lambda = 0.5$ so that both losses contribute equally. For vanilla training (without distillation), we set $\lambda = 0$, reducing the loss to standard masked cross-entropy. We use the AdamW optimizer with an initial learning rate of 10^{-3}, and batch sizes 512, 128, 64 for small, medium, and large models, respectively. During distillation, the teacher ESM-1b weights are frozen to provide logits for the KD loss; the student is updated via backprop on the combined loss.

4 Experiments and Results

Table 1 lists train and test set perplexities (PPLs) for all 17 model variants. The n-gram baselines are the worst-performing, with the best 4-gram achieving test PPL of 17.76. The student models perform considerably better, with the best 1.2M parameter student model having a test PPL of 14.21. The 33.7M parameter model achieves the lowest test PPL of 13.30 across all model variants and represents a 19.3x compression of ESM-1b, with a ΔTest PPL of +8.60. We find that while distillation helps for the 1.2M and 33.7M parameter models, the difference in performance to the vanilla student models is incremental, with the 5.6M parameter vanilla model outperforming its distilled versions. *We hypothesize that the true benefit of distillation emerges in data-scarce regimes, where the teacher's soft targets provide richer supervision than the raw labels alone.* In our case, the 50k training set may already provide sufficient signal for effective student generalization, especially for larger models. Furthermore, we observe that higher temperatures ($T > 1.0$), which induce smoother teacher distributions, tend to yield lower student perplexities, suggesting that softened targets can help prevent overfitting by encouraging better calibration and generalization.

Model	**Params (M)**	**Compression Rate ($\times$)**	**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**

Table 1: Final train and test perplexities for all models, grouped by parameter count. Distilled models are trained with different temperatures. Compression is relative to the 669M parameter ESM-1b baseline. ΔTest PPL indicates how much worse the model performs compared to ESM-1b.

4.1 Representation Evaluation

We perform a t-SNE clustering of the model's protein sequence representations to see if the student's embedding space clusters sequences by family, similar to ESM-1b. We prepend protein sequences with a special [CLS] token. Thus, we obtain the model's [CLS] token representation for each protein and use it to perform the t-SNE. The plots in Figure 1 show a 2D visualization of protein embeddings colored by their labels, i.e. Pfam or Superfamily, restricted to the top 10 most common label classes. Each point represents a protein, and the spatial grouping reflects how well the model clusters proteins with the same label in embedding space: tight, separated clusters suggest good semantic separation by the model. We show the t-SNE plots for ESM-1b, Student-h384-L8-VANILLA and Student-h384-L8-2.0 for both Pfam and Superfamily. Qualitatively, the distilled model's t-SNE protein representation clusters are more discernible/spread out relative to the teacher model.

Pfam is a well-known database of protein families defined by hidden Markov models [7], while structural superfamilies (as in the SCOPe database) group proteins by fold similarity [8]. Following ESM-1b [1], we use k-nearest neighbors (kNN) classification in embedding space: given a model's sequence embedding, we assign Pfam/Superfamily by majority vote among the nearest neighbors in a labeled training set. We extract each model's [CLS] embedding and perform k-NN classification on labeled test data. Tables 2 and 3 report accuracy, precision, recall and F1 (micro-averaged) for each of the 17 models on the Pfam and SCOPe Superfamily tasks, respectively. For the Pfam task, ESM-1b achieves the highest classification metrics. However, the 28-layer student model, distilled with a temperature of $T = 2.0$ only differs marginally. We see that distillation often yields improvements in classification metrics for all models. Notably, the distilled 28-layer student model achieves a +9$\%$ boost in accuracy, precision, and recall and a +10$\%$ boost in F1.

On the SCOPe superfamily classification task, the distilled 28-layer student model with temperature $T = 2.0$ beats all model variants, including the teacher. As before, there is a noticeable increase in classification metrics when distilling the 28-layer model with ESM-1b. Our analysis leads to a particularly noteworthy finding: *while distillation doesn't noticeably drive down perplexity for the large 33.7M parameter 28-layer student model, it substantially improves downstream protein annotation (Pfam/Supfam) classification, which it has not been explicitly trained to do.*

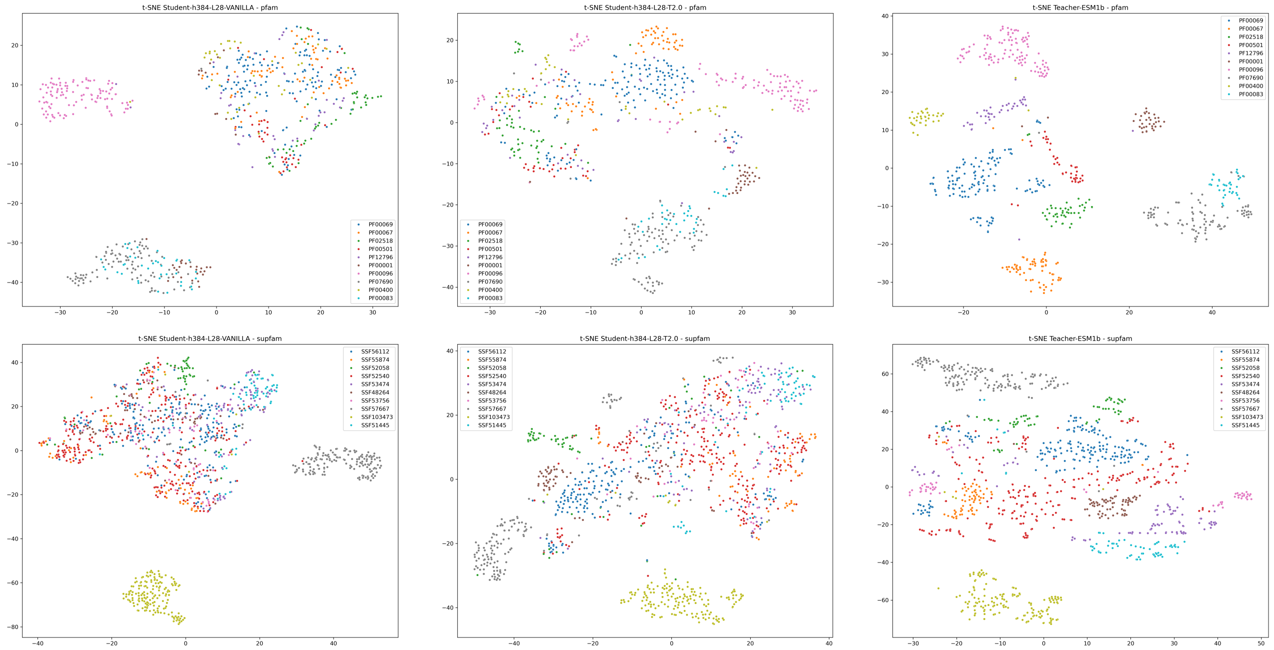

Figure 1: t-SNE plots of CLS token embeddings for Pfam (top row) and Superfamily (bottom row) labels. Each column corresponds to a different model: Student-h384-L8-VANILLA (left), Student-h384-L8-T=2.0 (middle), and ESM-1b (right). Each point represents a protein, colored by its label.

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 2: 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 3: Superfamily classification performance.

4.2 The Data-Scarce Regime

To test our hypothesis that knowledge distillation is particularly useful in the low data regime, we restrict ourselves to 1k training examples, 1k validation examples, and 1k test examples. Table 4 in Appendix A shows the train and test perplexities for all models. The 5.6M parameter model distilled with temperature $T = 1.0$ achieves a considerable reduction of -3.03 test PPL over its vanilla counterpart. For downstream k-NN protein annotation classification, we see a corresponding increase in F1 from 0.54 to 0.69 for Pfam and an increase from 0.50 to 0.57 for Superfamily. We also observe a -0.22 decrease in test PPL for the $T = 1.0$ distilled 1.2M parameter model over the vanilla baseline. For the 33.7M parameter models, it appears the perplexity is saturated and, hence, there is no benefit from distillation.

5 Conclusion

We show that ESM-1b can be effectively distilled into compact protein language models with up to 33.7M parameters, achieving strong language modeling performance and, in some cases, improved downstream protein annotation classification. While perplexity gains from distillation are modest, we find significant improvements in Pfam and SCOPe annotation tasks, especially for the largest student. Notably, the distilled 28-layer student outperforms the teacher in structural classification, despite being 20x smaller. This suggests that distillation not only compresses models, but can enhance biologically meaningful representations. We also find that in data scarce and low-capacity student model regimes, distillation can lead to a significant reduction in PPL.

References

[1] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[2] UniProt Consortium. The universal protein resource (uniprot) 2009. *Nucleic acids research*, 37(suppl_1):D169–D174, 2009.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[5] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[6] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[7] Erik LL Sonnhammer, Sean R Eddy, and Richard Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, 28(3):405–420, 1997.

[8] Loredana Lo Conte, Bart Ailey, Tim JP Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. Scop: a structural classification of proteins database. *Nucleic acids research*, 28(1):257–259, 2000.

6 Appendix A

Table 4 shows the train and test perplexities for all models in the data-scarce regime of 1k training examples. The best-in-class model variant is the 28-layer 33.7M parameter vanilla student model. The 5.6M parameter model distilled with temperature $T = 1.0$ achieves a considerable reduction of -3.03 test PPL over its vanilla counterpart. We see a -0.22 decrease in test PPL for the $T = 1.0$ distilled 1.2M parameter model over the vanilla baseline. For the 33.7M parameter models, it appears the perplexity is saturated and, hence, there is no benefit from distillation.

Model	Params (M)	Compression Rate	Train PPL	Test PPL	ΔTest PPL
ESM-1b	669.2	1.0	1.30	1.30	0.00
Student-h128-L8-VANILLA	1.2	557.7	19.63	19.65	18.35
Student-h128-L8-T0.5	1.2	557.7	19.44	19.48	18.18
Student-h128-L8-T1.0	1.2	557.7	19.46	**19.43**	**18.13**
Student-h128-L8-T2.0	1.2	557.7	19.60	19.59	18.29
Student-h256-L10-VANILLA	5.6	119.5	17.53	17.55	16.25
Student-h256-L10-T0.5	5.6	119.5	14.52	14.54	13.24
Student-h256-L10-T1.0	5.6	119.5	14.47	**14.52**	**13.22**
Student-h256-L10-T2.0	5.6	119.5	17.44	17.48	16.18
Student-h384-L28-VANILLA	33.7	19.9	14.41	**14.31**	**13.01**
Student-h384-L28-T0.5	33.7	19.9	14.57	14.57	13.27
Student-h384-L28-T1.0	33.7	19.9	14.54	14.59	13.29
Student-h384-L28-T2.0	33.7	19.9	16.62	16.61	15.31

Table 4: Train and test perplexities for all models, grouped by parameter count. ΔTest PPL is the increase in perplexity relative to ESM-1b's test PPL of 1.30.

Tables 5 and 6 show the metrics in the data-scarce regime for Pfam and SCOPe superfamily annotation classification, respectively.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
ESM-1b	**0.97**	**0.97**	**0.97**	**0.97**
Student-h128-L8-VANILLA	0.57	0.61	0.57	0.56
Student-h128-L8-T0.5	0.66	0.66	0.66	0.65
Student-h128-L8-T1.0	0.57	0.63	0.57	0.57
Student-h128-L8-T2.0	0.67	0.69	0.67	0.67
Student-h256-L10-VANILLA	0.55	0.57	0.55	0.54
Student-h256-L10-T0.5	0.69	0.71	0.69	0.69
Student-h256-L10-T1.0	0.66	0.66	0.66	0.65
Student-h256-L10-T2.0	0.42	0.39	0.42	0.40
Student-h384-L28-VANILLA	0.72	0.72	0.72	0.71
Student-h384-L28-T0.5	0.65	0.64	0.65	0.64
Student-h384-L28-T1.0	0.54	0.54	0.54	0.53
Student-h384-L28-T2.0	0.46	0.48	0.46	0.46

Table 5: Pfam classification performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
ESM-1b	**0.94**	**0.95**	**0.94**	**0.94**
Student-h128-L8-VANILLA	0.55	0.55	0.55	0.54
Student-h128-L8-T0.5	0.53	0.53	0.53	0.52
Student-h128-L8-T1.0	0.53	0.53	0.53	0.51
Student-h128-L8-T2.0	0.59	0.60	0.59	0.59
Student-h256-L10-VANILLA	0.51	0.52	0.51	0.50
Student-h256-L10-T0.5	0.58	0.58	0.58	0.57
Student-h256-L10-T1.0	0.58	0.58	0.58	0.57
Student-h256-L10-T2.0	0.35	0.35	0.35	0.34
Student-h384-L28-VANILLA	0.60	0.60	0.60	0.59
Student-h384-L28-T0.5	0.55	0.55	0.54	0.54
Student-h384-L28-T1.0	0.47	0.44	0.47	0.45
Student-h384-L28-T2.0	0.42	0.42	0.42	0.41

Table 6: Superfamily classification performance.