BIO-DISTILLER: BOOSTING SUPERVISED BASELINES BY DISTILLING BIOLOGICAL FOUNDATION MODELS

Mohan Vamsi Nallapareddy, Maria Boulougouri, Pierre Vandergheynst & Francesco Craighero*
Signal Processing Laboratory 2 (LTS2)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne
francesco.craighero@epfl.ch

ABSTRACT

The success of foundation models has driven their application to biological data, including RNA, DNA, and protein sequences. These are expected to adapt to new tasks, while also providing novel insights into molecular function and biological mechanisms underlying health and disease. Despite recent advances, benchmarks show that biological foundation models fail to consistently outperform simpler supervised approaches. Moreover, key challenges remain, such as the development of faithful interpretation methods and the integration of multiple modalities. Here, we propose BIO-Distiller, a framework to distill the rich information from biological foundation models into smaller models. In our benchmark on six RNA downstream tasks, we first show that well-tuned supervised baselines can still outperform foundation models. Furthermore, knowledge distillation consistently boosts the baselines' performance by up to 10% across all tasks. Additionally, our framework is capable of integrating multiple foundation models, whether from the same modality by exploiting design and pre-training differences, or across different modalities. The results not only confirm the potential of BIO-Distiller but also provide guidelines for its application to new tasks and modalities, paving the way toward high-performing, efficient, and easily interpretable supervised models for biology.

1 INTRODUCTION

The impressive performance of Foundation Models (FMs) Bommasani et al. (2022) in domains such as Natural Language Processing and Computer Vision has demonstrated their exceptional capacity to exploit large-scale pre-training to generalize across diverse tasks and modalities. This success has prompted their adoption in the biomedical domain Liu et al. (2025), including the introduction of biological FMs Consens et al. (2025); Benegas et al. (2025), pre-trained on RNA Chen et al. (2022); Zhang et al. (2023); Chu et al. (2024), DNA Nguyen et al. (2023), and protein sequences Lin et al. (2023).

Despite significant advances, biological FMs still face critical challenges that limit their practical applicability Benegas et al. (2025); Sapoval et al. (2022). Benchmark studies have shown that neither fine-tuning FMs Xu et al. (2024) nor utilizing their frozen embeddings Marin et al. (2023); Tang et al. (2025) consistently surpasses traditional supervised methods on genomic downstream tasks. Additionally, most biological FMs remain unimodal, with only recent efforts focusing on multimodal integration techniques Garau-Luis et al. (2025) and multimodal pre-trained models He et al. (2025). Finally, interpretability is still a major open issue, as existing approaches mainly depend on post hoc analyses that are often constrained by architectural design choices such as the chosen tokenization strategy Chen et al. (2024) and the input modality, which cannot be adapted to the downstream tasks.

In this work, we explore Knowledge Distillation (KD) Hinton et al. (2015) as a promising approach to tackle the limitations of biological FMs. Originally developed to transfer knowledge from large

*Corresponding author

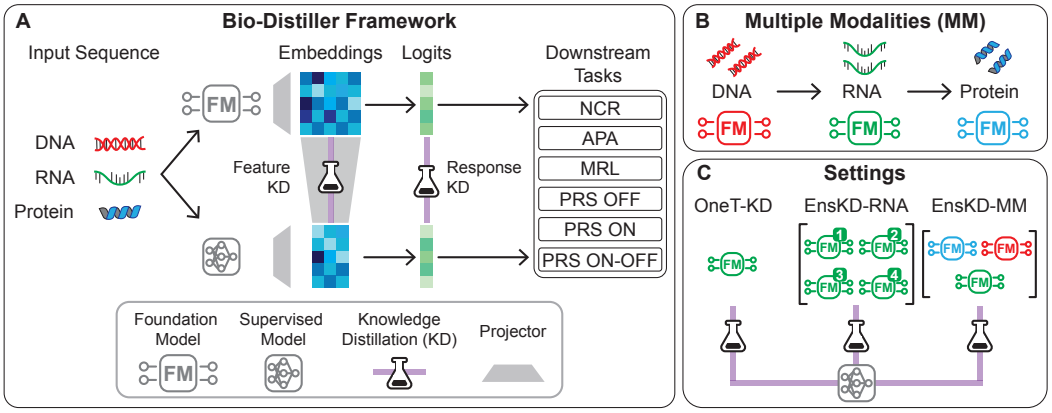


Figure 1: **Overview of the BIO-Distiller framework.** (**A**) The framework can take RNA, DNA, or protein sequences as input. It conducts Knowledge Distillation (KD) using information from the Foundation Models (FMs) pre-trained on the input sequence modality. It performs KD either using the representations (Feature KD) or using the logits (Response KD) of the teacher FM. These models are trained to predict six RNA-related downstream tasks. (**B**) FMs across modalities. (**C**) Settings: One Teacher KD (One-T KD), Ensemble KD across RNA-based (EnsKD-RNA) and multimodal FMs (EnsKD-MM).

models or ensembles to smaller, more efficient ones, KD allows these compact models to maintain high performance while substantially improving inference efficiency. Little work exists on KD applications in genomics, and only recently has it been used to distill ensembles into student models that preserve generalization performance Avsec et al. (2026), along with their increased robustness Zhou et al. (2026).

In this study, we explore KD as a novel and versatile strategy to harness the strengths of pre-trained biological FMs, while overcoming some of their major applicability limitations. To that end, we introduce the BIO-Distiller framework to enable the distillation of knowledge from one or more biological FMs into a supervised model of our choice. We address the lack of existing guidelines by performing an extensive evaluation of diverse distillation approaches across six RNA sequence-level downstream tasks from an existing benchmark Ren et al. (2025) and six different biological FMs across multiple modalities. Moreover, we re-imagine multi-teacher distillation as a novel paradigm to integrate biological FMs across different modalities. In the current version, we test a well-tuned standard Convolutional Neural Network as the student model. However, considering that our framework is designed to be model-agnostic, it can be extended to dataset-specific alternatives, addressing the design constraints of adapting FMs in downstream tasks.

Contributions.

- We validate prior results on DNA tasks Xu et al. (2024) by showing that well-tuned supervised baselines can perform comparably or even outperform RNA-based FMs on RNA-related downstream tasks.
- We show that distilling from a single RNA-based FM can boost the performance of the supervised baseline by up to 10%.
- Given our extensive evaluation of diverse distillation strategies, we provide the first guidelines to apply knowledge distillation techniques on genomic tasks.
- We demonstrate that a simple average ensemble of RNA-based FMs, exploiting their distinct design choices, can consistently result in the highest performance across tasks.
- We propose knowledge distillation as a novel method to integrate FMs across different biological modalities.
- We confirmed the ability of our distilled supervised models to extract relevant biological insights using a standard interpretability pipeline.

2 BACKGROUND

2.1 BIOLOGICAL FOUNDATION MODELS

Foundation Models (FMs) are ML models pre-trained on large-scale datasets, allowing their adaptation to a wide range of downstream tasks Bommasani et al. (2022). Recently, there has been growing interest in developing such models tailored for biological domains, such as RNA, DNA, and protein sequences Liu et al. (2025); Consens et al. (2025); Benegas et al. (2025).

Despite significant interest in developing biological FMs, several challenges hinder their effectiveness on downstream tasks. Pre-training strategies that have delivered strong results in Natural Language Processing often fail to translate to the genomics domain. Notably, supervised baselines Xu et al. (2024); Marin et al. (2023); Tang et al. (2025) and even randomly initialized genomic FMs Vishniakov et al. (2024) have demonstrated comparable performance to pre-trained models, calling into question the practical benefits of current pre-training approaches in genomics. Moreover, a key objective of applying deep learning to biology is not only to learn complex interactions and patterns from input data, but also to extract meaningful biological insights from these models Talukder et al. (2021); Sapoval et al. (2022); Novakovsky et al. (2023). Although attention weights and post-hoc methods like SHAP Lundberg & Lee (2017) offer ways to interpret input features, uncovering global patterns and decision rules, particularly in large-scale models like FMs, remains an active and unresolved area of research Bereska & Gavves (2024). Finally, many biological processes, such as isoform expression and translation efficiency, depend on the complex interplay between DNA, RNA, and protein sequences. To make more accurate predictions, models must integrate information across all these modalities. However, most biological FMs developed to date have been unimodal. While recent studies have explored methods for combining embeddings from different biological FMs Garau-Luis et al. (2025) and pre-training large-scale multimodal models He et al. (2025), this still remains an active and evolving area of research.

2.2 KNOWLEDGE DISTILLATION

The need for powerful and high-performing FMs in conjunction with the need to design smaller, efficient models has rapidly increased interest in the domain of knowledge distillation Yang et al. (2024); Sun et al. (2023); Vemulapalli et al. (2024). Distillation was originally proposed as a way to transfer information from a larger ensemble of models into a smaller student model by providing it "soft-labels" from the teacher to serve as additional guidance Hinton et al. (2015). When the student model is an easily interpretable one, such as a decision tree, distillation can then be used to explain the decision of the larger black-box teacher Frosst & Hinton (2017). More broadly, these techniques have been primarily employed in domains such as computer vision and natural language processing Gou et al. (2021), but there is little work studying this in biological domains. Here, recent work showcases a novel ensemble distillation framework, titled DEGU, to distill both the output and its epistemic uncertainty estimate from the teacher ensemble into the student model Zhou et al. (2026). Similarly, this approach has been employed in the AlphaGenome framework to compress a high-performing ensemble of supervised models Avsec et al. (2026).

3 METHODS

In this section, we outline the BIO-Distiller framework (ref fig. 1) for conducting knowledge distillation on biological FMs. We describe the different components of it, including the different FMs, supervised baseline, downstream tasks, knowledge distillation approaches, and distillation settings.

3.1 SUPERVISED BASELINE AND STUDENT MODEL

We use a simple three-layer Convolutional Neural Network (CNN) both as a baseline supervised model and as a student model for evaluating knowledge distillation, following an architecture design similar to prior work in genomics Grešová et al. (2023); Zhou & Troyanskaya (2015). Each convolutional block consists of a one-dimensional convolution, followed by batch normalization, ReLU activation, max pooling with a kernel size ks, a stride of 2, and dropout with probability p_{drop}. The initial number of channels, denoted by ch_{init}, are halved after each layer. After the final

convolutional block, global average pooling is applied over the sequence length, and the resulting feature vector is passed through two fully connected layers of equal size, producing the final logits.

3.2 Biological Foundation Models

We employed four RNA-based FMs: RNA-FM Chen et al. (2022), TE+EL and MRL variants of UTR-LM Chu et al. (2024), and RNA-MSM Zhang et al. (2023). Both RNA-FM and RNA-MSM are BERT-like Devlin et al. (2019) models pre-trained on raw sequences and multiple sequence alignments of non-coding RNA sequences, respectively. Whereas the UTR-LM models utilize a transformer-based architecture Vaswani et al. (2017) and were pre-trained on sequences, secondary structures, and minimum free energies of RNA 5' untranslated regions. These UTR-LM models were further fine-tuned on the Mean Ribosome Loading task to obtain the MRL variant, and translation efficiency and mRNA expression level tasks to obtain the TE+EL variant. Furthermore, for DNA and protein sequences, we used the HyenaDNA (tiny-1k) Nguyen et al. (2023), and ESM-2 (t6-8M) Lin et al. (2023) models, respectively.

We define the ensemble model of n FMs (EnsAvg) as the average of their output logits ($\boldsymbol{l}_i$):

$$l_{\text{EnsAvg}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{l}_i \tag{1}$$

We will consider two ensemble types, one composed of all four RNA-based FMs (EnsAvg-RNA) and the other composed of one FM from each of the three biological modalities (EnsAvg-MM).

3.3 Downstream Tasks

We evaluate our approach on six tasks from the BEACON benchmark Ren et al. (2025), each targeting different RNA properties. Five tasks are formulated as regression problems: (1) Mean Ribosome Load (MRL) Sample et al. (2019), which quantifies translation efficiency by predicting ribosome density, (2) Alternative Polyadenylation Isoform Prediction (APA) Bogard et al. (2019), which models isoform expression driven by alternative polyadenylation, a mechanism that generates mRNA variants by choosing different cleavage sites, and (3–5) Programmable RNA Switch Activities (PRS ON, PRS OFF, PRS ON-OFF) Angenent-Mari et al. (2020), which model the response of synthetic RNA switches to external signals across three different modes, activation (ON), repression (OFF), and dual (ON-OFF). The sixth task is a multiclass classification problem: (6) Non-coding RNA Function (NCR) Amin et al. (2019), which assigns RNA sequences to functional categories (e.g., transfer RNA, micro RNA, ribosomal RNA).

3.4 Losses and Metrics

To accommodate different task types, we use task-specific losses denoted by $\mathcal{L}_{\text{task}}$. For regression tasks, we apply the Mean Squared Error (MSE) loss, while for the classification task, we use the Categorical Cross-Entropy (CCE) loss.

To transfer the knowledge from a teacher to a student model, we introduce a knowledge distillation loss $\mathcal{L}_{\text{KD}}$. Following the taxonomy in Gou et al. (2021), we apply either feature-based distillation, by aligning the hidden representations of the teacher and student, or response-based distillation, by encouraging the student's logits to match those of the teacher (see fig. 1).

Let $\boldsymbol{t} \in \mathbb{R}^n$ and $\boldsymbol{s} \in \mathbb{R}^m$ denote the activation vectors extracted from the penultimate layers of the teacher and student models, respectively. Given a learnable projection matrix $\boldsymbol{W}^{\text{proj}} \in \mathbb{R}^{n \times m}$ that maps the student's output to the teacher's feature space, we define the projected student representation as $\boldsymbol{s}^{\text{proj}} = \boldsymbol{W}^{\text{proj}} \boldsymbol{s}$. To distill the feature-based knowledge of the teacher by matching $\boldsymbol{t}$ and $\boldsymbol{s}^{\text{proj}}$, we either use the MSE as in FitNets Romero et al. (2015):

$$\mathcal{L}_{\text{fitnet}} = \frac{1}{n} \left\| \boldsymbol{t} - \boldsymbol{s}^{\text{proj}} \right\|_2^2 \tag{2}$$

or a combination of the Euclidean distance and the cosine similarity (referred to as "eucosine"):

$$\mathcal{L}_{\text{eucosine}} = \frac{1}{n} \left\| \boldsymbol{t} - \boldsymbol{s}^{\text{proj}} \right\|_2 + \left(1 - \frac{\boldsymbol{t}^\top \boldsymbol{s}^{\text{proj}}}{\|\boldsymbol{t}\| \, \|\boldsymbol{s}^{\text{proj}}\|} \right) \tag{3}$$

To distill the response-based knowledge of the teacher, we use the "logit" loss, which is defined as

$$\mathcal{L}_{\text{logit}} = \frac{1}{p} \left\| \boldsymbol{l}^t - \boldsymbol{l}^s \right\|_2^2 \tag{4}$$

where $\boldsymbol{l}^t$ and $\boldsymbol{l}^s$ are the p-dimensional logits of the teacher and the student, respectively.

The final student loss is a linear combination of the task loss and one of the three KD losses:

$$\mathcal{L}_{\text{student}} = \alpha \mathcal{L}_{\text{task}} + \beta \mathcal{L}_{\text{KD}} \tag{5}$$

We considered two cases: "sum" with $\alpha = \beta = 1$ and "convex" with $\alpha = (1 - \alpha_{KD})$ and $\beta = \alpha_{KD}$, where α_{KD} is designed to be a tunable hyperparameter.

To support multiple teachers (k in total), we modify both the feature-based and response-based KD protocols. For feature-based KD, we concatenate the penultimate-layer activations of the teachers, that is $\boldsymbol{t} = [\boldsymbol{t}^1; \boldsymbol{t}^2; \ldots; \boldsymbol{t}^k]$, where $\boldsymbol{t}^i$ is the activation of the i-th teacher. For the response-based KD, we compute the average of the teachers' logits, that is $\boldsymbol{l} = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{l}_i$ where $\boldsymbol{l}_i$ represents the logits of the i-th teacher. The corresponding student models are then trained to match the concatenated activations (for feature KD) or the averaged logits (for response KD), depending on the distillation strategy.

To evaluate the model performances, we considered the R^2 score, Pearson Correlation Coefficient (PCC), and MSE for the regression tasks, and Weighted F1 (WF1), Balanced Accuracy Score (BAS), and Mathews Correlation Coefficient (MCC) for the classification task.

3.5 Hyperparameter Tuning

All the baselines and student models were tuned using the Asynchronous Successive Halving Algorithm (ASHA) Li et al. (2020), as implemented in the Ray Tune library Liaw et al. (2018). The search space is detailed in table A.2. We configured Ray Tune to evaluate 100 hyperparameter configurations, with ASHA using a reduction factor of 2 and a grace period of 10 epochs. The Adam optimizer Kingma & Ba (2017) was used, with the learning rate reduced by 10% if validation performance did not improve for 5 epochs. Each trial ran for up to 100 epochs, with early stopping based on a patience of 10 epochs.

For the FMs, we tuned a smaller set of hyperparameters using a grid search. The tested values are listed in table A.2. We used the AdamW optimizer Loshchilov & Hutter (2018) in conjunction with a cosine annealing scheduler with a minimum learning rate of 10^{-6}. Training was conducted for a maximum of 100 epochs, with early stopping based on a patience of 20 epochs.

3.6 BIO-Distiller Settings

The BIO-Distiller framework enables knowledge distillation from one or more teacher biological FMs into student CNNs. This framework can be employed in three different settings: "OneT-KD", "EnsKD-RNA", and "EnsKD-MM". In OneT-KD, we conduct distillation using a single RNA-based FM out of the set of four, with or without fine-tuning on the downstream task. For EnsKD-RNA, we use an ensemble of the four fine-tuned RNA-based FMs, whereas for EnsKD-MM, we use one RNA-based FM, HyenaDNA, and ESM-2. In each setting, we explore different knowledge distillation variants by testing multiple KD losses (fitnet, eucosine, and logits), loss aggregation methods (sum and convex), and FMs, either fine-tuned or frozen.

3.7 Interpretability

To assess the reliability of our supervised baseline beyond predictive performance, we applied a standard genomics interpretability pipeline to extract relevant RNA motif sequences. We conduct this analysis on the Alternative Polyadenylation Isoform Prediction (APA) task. Employing the tangermeme library Schreiber (2025) on a given pre-trained model, we first extract feature attributions via Deep SHAP Lundberg & Lee (2017)[1] for 10 000 randomly selected test sequences,

[1]Reference generated with the default 20 dinucleotide shuffles.

and identify high-attribution windows, known as seqlets[2]. These seqlets were then annotated with protein-binding RNA motifs from a published compendium Ray et al. (2013) using TomTom Gupta et al. (2007). The resulting motifs were functionally enriched via Enrichr Kuleshov et al. (2016) through the GSEApy Fang et al. (2023) library using the following gene sets: GO Biological Process Ashburner et al. (2000), KEGG Kanehisa et al. (2025), and Reactome pathways Milacic et al. (2024).

4 RESULTS

4.1 BENCHMARKING DISTILLATION FROM RNA-BASED FMS

In our initial benchmark, we evaluated four RNA-based FMs (RNA-FM, RNA-MSM, UTR-LM (MRL), and UTR-LM (TE+EL)) alongside our supervised CNN baseline (referred to as "Baseline") across six distinct RNA downstream tasks (table 1). Consistent with prior findings on these tasks and models Ren et al. (2025), no single FM consistently outperforms the others. In our experiments, UTR-LM (TE+EL) achieved the best performance on PRS tasks and MRL, while RNA-FM led in NCR, and UTR-LM (MRL) performed the best on APA. We notice that FMs such as RNA-FM and RNA-MSM, which have been pre-trained on non-coding RNA sequences, achieve greater performance on the NCR task. Conversely, UTR-LM (MRL) and UTR-LM (TE+EL), which have been pre-trained on translation efficiency related tasks, demonstrate higher performance on the MRL task. These findings highlight the importance of different pre-training strategies and their effect on specific downstream applications.

Metric	MCC	R^2				
Task / Model	NCR	APA	MRL	OFF	ON	ONF
RNA FMs						
RNA-FM	**96.38**	86.73	<u>82.42</u>	35.18	67.13	33.12
RNA-MSM	90.73	88.21	79.57	31.71	66.55	36.39
UTR-LM (MRL)	86.73	**88.81**	82.05	47.11	73.10	44.42
UTR-LM (TE+EL)	82.64	88.53	**82.94**	49.76	73.26	45.50
CNNs						
Baseline	84.15	87.74	74.80	50.26	<u>74.14</u>	45.73
Best OneT-KD	<u>94.64</u>	<u>88.80</u>	77.24	**51.26**	**74.85**	**47.78**

Table 1: **Single teacher KD performance**. Performance of RNA-based FMs and baseline against the best-performing distilled baseline with a single RNA-based FM as the teacher (OneT-KD). PRS ON, PRS OFF, and PRS ON-OFF are mentioned as ON, OFF, and ONF, respectively. Configurations of the Best OneT-KD models are mentioned in table A.1, and additional metrics are provided in the Appendix.

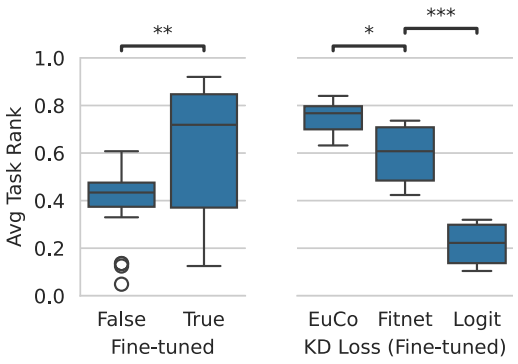
Through extensive hyperparameter optimization, the baseline model attains high performance on all the downstream tasks, outperforming or performing comparably to FMs on four out of six tasks. Our results in RNA tasks align with conclusions from previous studies on DNA tasks Xu et al. (2024), showing that well-tuned supervised models can challenge the performance of pre-trained FMs.

Using our BIO-Distiller framework, we trained several variants of the baseline model, each incorporating knowledge distilled from a single RNA-based FM. Each variant used a different FM as the teacher, either fine-tuned or kept frozen, and applied one of the three available knowledge distillation loss functions along with one of two possible loss aggregation strategies. The best variant in each task (table A.1), referred to as "Best OneT-KD" (Best One Teacher KD), consistently boosts the performance of the baseline model on all six tasks. Performance gains reach around 2% for regression tasks, while the NCR classification task showcases a notable performance boost of 10% (table 1). These results highlight BIO-Distiller's effectiveness in transferring knowledge from RNA-based FMs to simpler models, significantly narrowing the performance gap between them.

4.2 COMPARING KNOWLEDGE DISTILLATION VARIANTS

Using our BIO-Distiller framework, we evaluated multiple knowledge distillation variants across different FMs, KD losses, and loss aggregation methods, allowing us to identify the configuration that consistently delivers the best results.

[2]Seqlets were extracted using `recursive_seqlets()`.



(a)

Metric	MCC	R^2				
Task	NCR	APA	MRL	OFF	ON	ONF
Integrating multimodal FMs						
Best RNA	**96.38**	88.81	82.94	49.76	73.26	45.50
HyenaDNA	82.39	81.98	67.01	38.88	61.81	31.58
ESM-2	77.28	67.75	31.74	36.85	65.57	32.37
EnsAvg-MM	94.43	84.84	74.55	48.68	71.94	43.43
EnsKD-MM	93.31	88.46	73.00	50.80	74.36	46.45
Ensembling RNA FMs						
EnsAvg-RNA	94.46	**90.21**	**85.10**	**51.84**	**75.31**	**48.91**
EnsKD-RNA	94.77	88.68	74.53	51.12	74.81	47.79

(b)

Figure 2: (a) **Rank comparisons between OneT-KD distillation variants.** We averaged performance percentile ranks over all OneT-KD variants, using R^2 for regression and MCC for classification. The rankings were grouped by (left) whether the teacher was fine-tuned and (right) the distillation loss $\mathcal{L}_{KD}$, considering only configurations with a fine-tuned teacher. A two-sided Mann–Whitney–Wilcoxon test was used to assess statistical significance (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$). (b) **Ensemble and multimodal integration**. Comparison of unimodal FMs for RNA (only the best one is reported), DNA, and protein sequence with their average ensemble (EnsAvg-MM) and the corresponding distilled baseline (EnsKD-MM). Likewise, comparison of the RNA-based FMs ensemble (EnsAvg-RNA) with its distilled baseline (EnsKD-RNA). PRS ON, PRS OFF, and PRS ON-OFF are mentioned as ON, OFF, and ONF, respectively. Additional metrics are provided in the Appendix.

Although fine-tuning the teacher model requires additional computational resources, we observed that having a fine-tuned teacher yields significantly better results (fig. 2a, left). Focusing on the top-performing variants with a fine-tuned teacher, the eucosine distillation loss yields significantly higher performance, followed by fitnet, while the logit loss performs the poorest (fig. 2a, right).

Consistent with the observation that no single FM excels across downstream tasks, we found no teacher FM outperformed the others during distillation. Moreover, in only half of the tasks did the best-performing FM also serve as the most effective teacher (table A.1). These findings suggest that, even in terms of the knowledge captured in the FM, there is still no universally superior model. With respect to the loss aggregation strategy, since the convex approach does not offer a significant advantage over simple summation (table A.1), the latter approach is preferred due to its simplicity and comparable effectiveness.

4.3 RNA FMs Ensembles and Multimodal Integration

RNA-based FMs differ in their architectures, pre-training strategies, and source datasets. Since no single model consistently outperforms the others across all tasks, each FM tends to excel in different areas. This becomes especially evident when we combine the predictions of our four RNA-based FMs through simple averaging (EnsAvg-RNA), forming an ensemble that achieves the highest overall performance across all tasks, with the only exclusion of NCR (fig. 2b). To leverage the diversity in their learned representations, we define a new approach (EnsKD-RNA) in which we employ all four FMs collectively as an ensemble teacher. Since fine-tuning the teacher model combined with the eucosine loss produced the best results, we focused exclusively on this setup, along with the use of the simpler sum loss aggregation. The EnsKD-RNA model for the NCR classification task outperforms EnsAvg-RNA and OneT-KD approaches, whereas it attains comparable performances with the latter approaches for the other regression tasks.

Given that our proposed ensemble teacher approach is modality-agnostic, we used it to distill combinations of FMs from different modalities into our supervised baseline. Considering our focus on RNA downstream tasks, integrating DNA, RNA, and protein models was a logical choice, as these complementary modalities capture interconnected layers of biological function. To assess

their respective contribution, we first test one DNA FM (HyenaDNA) and one protein sequence FM (ESM-2) across all tasks and compare them with the best-performing RNA-based FM. We observe that among the three, the RNA model consistently performs best, followed by HyenaDNA, which is better than ESM-2 on four out of six tasks. This could indicate that RNA is the most informative modality for these downstream tasks, followed by DNA and then protein sequences.

We apply our ensembling strategies to these three multimodal FMs through the EnsAvg-MM and EnsKD-MM models, with EnsKD-MM introducing a novel approach to integrating multimodal FMs. Across our tasks, EnsAvg-MM does not attain performances comparable to those of the best-performing constituent FM, showcasing the need for alternative approaches to aggregate these modalities given their inconsistent performances. The EnsKD-MM model takes into account the same setup as EnsKD-RNA and surpasses EnsAvg-MM on four out of six tasks. Moreover, it outperforms each individual FM on all three PRS-related tasks and achieves comparable performance to the top-performing FM on the remaining three tasks.

4.4 Interpretability

To evaluate whether our simple supervised model not only predicts target variables accurately but also captures biologically meaningful input features, we apply our pipeline for extracting genomic insights on the best distilled baseline with a single teacher ("OneT-KD", table 1). We center our analysis on the Alternative Polyadenylation Isoform Prediction (APA) task, as it enables a simple assessment of whether the model attends to sequence regions involved in RNA regulation and alternative splicing. Both are key functions influencing isoform expression Jones et al. (2024), which is the target variable for this task.

Using the tangermeme library Schreiber (2025), we identified 91 unique known motifs in test sequence regions with high attribution scores, indicating where the model primarily focuses. In fig. 3A and B, we highlight two of these motifs: "KHDRBS1" and "HNRNPA1". Both of them ranked among the top-10 motifs with the highest attribution scores, and they have been observed to be well-established regulators of alternative splicing Jia et al. (2019); Frisone et al. (2015). Considering the full set of motifs, in fig. 3C we report the top-10 enriched functions of our motif set according to their Adjusted P-value (always $< 5 \times 10^{-22}$). This showcases that the model clearly captures regions in the input sequence with functions affecting isoform expression, with splicing being the most important.

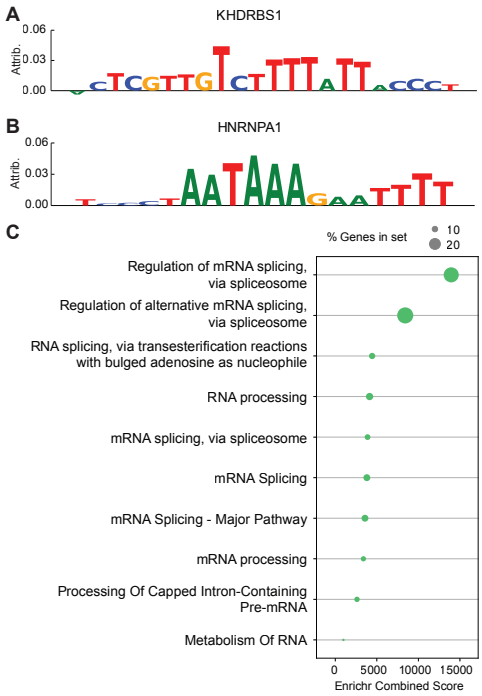


Figure 3: **Motif analysis on the best OneT-KD model on the APA task.** (**A & B**) report two of the top-10 motifs according to the model attributions, together with their annotation. (**C**) Over-representation analysis via Enrichr of all the identified motifs.

5 Discussion

In this work we introduce BIO-Distiller, a novel approach to exploit pre-trained biological Foundation Models (FMs) by distilling their knowledge into smaller supervised models.

In our tests performed across 6 RNA datasets from the BEACON benchmark Ren et al. (2025), we take into account four RNA-based FMs together with a well-tuned Convolutional Neural Network (CNN) baseline. Our initial results align with previous findings on DNA benchmarks Xu et al. (2024), showing that simple yet carefully optimized baselines can match the performance of specialized FMs. Through our BIO-Distiller framework, we combined the strengths of well-tuned CNNs with Knowledge Distillation (KD). Here, we evaluated multiple variants where each RNA-based FM

served as a teacher, both with and without fine-tuning, and tested different distillation approaches. Our results highlight how distillation can consistently improve the supervised baseline by up to 10%, often matching or even outperforming the performance of the FMs. Moreover, leveraging our extensive experiments, we were able to identify the distillation approaches that yield significant performance improvements, such as fine-tuning the teacher FM, offering valuable guidance for future applications.

Since most current biological FMs are unimodal, there is growing interest in integrating models across modalities Garau-Luis et al. (2025). Within our BIO-Distiller framework, we implemented and evaluated distillation as a promising strategy to combine knowledge from multiple models simultaneously, with all serving concurrently as teachers. Specifically, by including two additional FMs for DNA and protein sequences, we compared the performance achieved by aggregating one FM for each modality, against using all the RNA-based FMs together. Given a set of FMs, we considered both averaging their predictions and using them as teachers for our distilled baseline. While integrating multiple modalities should help boost the performance, especially due to the interconnectedness of DNA, RNA, and proteins, in our experiments, ensembling RNA-based FMs achieves higher performance, showcasing how even FMs from the same modality might provide complementary information. Regarding ensemble distillation, only in the non-coding RNA function classification task were we able to improve over the single teacher setting.

Lastly, it is important to assess the efficacy of these models beyond simple performance metrics. We exemplify this in the current work by adding a task-specific interpretability pipeline. By implementing a standard method to identify protein-binding RNA motifs in the OneT-KD model trained on the Alternative Polyadenylation Isoform Prediction (APA) task, we manage to extract sequences from the input that are functionally related to the prevalence of RNA isoforms. Specifically, our analysis outputs motifs that are significantly involved in the regulation of RNA processing, splicing, and regulation, which lead to the differential occurrence of alternative RNA isoforms. This validated the effectiveness of the simple supervised baseline beyond its performance evaluation.

6 Future Work

Among the six RNA benchmark tasks, non-coding RNA function prediction exhibited the largest performance gains from distillation, both with single and multiple teachers, and was also the only classification task. Since most existing distillation methods are designed keeping classification tasks in mind Zhou & Chiam (2023), future research on distillation should further prioritize regression tasks, which are more prevalent in real-world biological applications. Moreover, other future extensions of BIO-Distiller will include DNA benchmarks Marin et al. (2023); Grešová et al. (2023) and related FMs, but also alternative supervised baselines such as the UNet Ronneberger et al. (2015).

More broadly, BIO-Distiller holds significant promise for advancing interpretable deep learning in genomics without compromising model accuracy. In this study, we have primarily focused on performance gains and have only used interpretability analysis on one task as a validation. Future work will explore how the BIO-Distiller framework can enhance explanations, for example, by improving their robustness and faithfulness. This objective is especially feasible when the distilled model is interpretable by design, such as in the case of Self-Explaining Neural Networks Alvarez Melis & Jaakkola (2018), improving the way in which we extract genomic insights.

Acknowledgments

F.C. and M.V.N. are funded by a Swiss National Science Foundation (SNSF) Sinergia grant (CRSII5-205884). M.B. is funded by the Graph Neural Networks for Explainable Artificial Intelligence ERA-NET + EJP (20CH21-195579) grant.

Code and Extended Data

Code and extended data is available at `https://github.com/epfl-lts2/biod-mlgenx`.

REFERENCES

David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Noorul Amin, Annette McGrath, and Yi-Ping Phoebe Chen. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence*, 1(5):246–256, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0051-2.

Nicolaas M. Angenent-Mari, Alexander S. Garruss, Luis R. Soenksen, George Church, and James J. Collins. A deep learning approach to programmable RNA switches. *Nature Communications*, 11 (1):5057, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18677-1.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, January 2026.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. Genomic language models: Opportunities and challenges. *Trends in Genetics*, 41(4):286–302, April 2025. ISSN 0168-9525. doi: 10.1016/j.tig.2024.11.013.

Leonard Bereska and Stratis Gavves. Mechanistic Interpretability for AI Safety - A Review. *Transactions on Machine Learning Research*, April 2024. ISSN 2835-8856.

Nicholas Bogard, Johannes Linder, Alexander B. Rosenberg, and Georg Seelig. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell*, 178(1):91–106.e23, June 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.04.046.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions, August 2022.

Valerie Chen, Muyu Yang, Wenbo Cui, Joon Sik Kim, Ameet Talwalkar, and Jian Ma. Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nature Methods*, 21(8):1454–1461, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02359-7.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, 6(4):449–460, April 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00823-9.

Micaela E. Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J. Theis, Alan Moses, and Bo Wang. Transformers and genome language models. *Nature Machine Intelligence*, pp. 1–17, March 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01007-9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. GSEApy: A comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*, 39(1):btac757, January 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac757.

Paola Frisone, Davide Pradella, Anna Di Matteo, Elisa Belloni, Claudia Ghigna, and Maria Paola Paronetto. SAM68: Signal Transduction and RNA Metabolism in Human Cancer. *BioMed Research International*, 2015(1):528954, 2015. ISSN 2314-6141. doi: 10.1155/2015/528954.

Nicholas Frosst and Geoffrey Hinton. Distilling a Neural Network Into a Soft Decision Tree, November 2017.

Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Maša Roller, Bernardo de Almeida, Christopher Blum, Lorenz Hexemer, Stefan Laurent, Maren Lang, Thomas Pierrot, and Guillaume Richard. Multi-modal Transfer Learning between Biological Foundation Models. *Advances in Neural Information Processing Systems*, 37:78431–78450, January 2025.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819, June 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01453-z.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: A collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, May 2023. ISSN 2730-6844. doi: 10.1186/s12863-023-01123-8.

Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, February 2007. ISSN 1474-760X. doi: 10.1186/gb-2007-8-2-r24.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, Feng Zhu, Edward C. Holmes, Jieping Ye, Jun Li, Yuelong Shu, Mang Shi, and Zhaorong Li. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01044-4.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015.

Qi Jia, Hongbo Nie, Peng Yu, Baiyun Xie, Chenji Wang, Fu Yang, Gang Wei, and Ting Ni. HNRNPA1-mediated 3′ UTR length changes of *HN1* contributes to cancer- and senescence-associated phenotypes. *Aging*, 11(13):4407–4437, June 2019. ISSN 1945-4589. doi: 10.18632/aging.102060.

Emma F Jones, Anisha Haldar, Vishal H Oza, and Brittany N Lasseigne. Quantifying transcriptome diversity: A review. *Briefings in Functional Genomics*, 23(2):83–94, March 2024. ISSN 2041-2657. doi: 10.1093/bfgp/elad019.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: Biological systems database as a model of the real world. *Nucleic Acids Research*, 53 (D1):D672–D677, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae909.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, July 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw377.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A System for Massively Parallel Hyperparameter Tuning, March 2020.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

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, March 2023. doi: 10.1126/science.ade2574.

Xiangrui Liu, Yuanyuan Zhang, Yingzhou Lu, Changchang Yin, Xiaoling Hu, Xiaoou Liu, Lulu Chen, Sheng Wang, Alexander Rodriguez, Huaxiu Yao, Yezhou Yang, Ping Zhang, Jintai Chen, Tianfan Fu, and Xiao Wang. Biomedical Foundation Model: A Survey, March 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, September 2018.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks. In *The Twelfth International Conference on Learning Representations*, October 2023.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1025.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, December 2023.

Gherman Novakovsky, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, February 2023. ISSN 1471-0064. doi: 10.1038/s41576-022-00532-2.

Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, July 2013. ISSN 1476-4687. doi: 10.1038/nature12311.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, Dong Yuan, Wanli Ouyang, and Xihui Liu. BEACON: Benchmark for Comprehensive RNA Tasks and Language Models. *Advances in Neural Information Processing Systems*, 37:92891–92921, January 2025.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets, March 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Paul J. Sample, Ban Wang, David W. Reid, Vlad Presnyak, Iain J. McFadyen, David R. Morris, and Georg Seelig. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology*, 37(7):803–809, July 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0164-5.

Nicolae Sapoval, Amirali Aghazadeh, Michael G. Nute, Dinler A. Antunes, Advait Balaji, Richard Baraniuk, C. J. Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, R. A. Leo Elworth, Bryce Kille, Anastasios Kyrillidis, Luay Nakhleh, Cameron R. Wolfe, Zhi Yan, Vicky Yao, and Todd J. Treangen. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29268-7.

Jacob Schreiber. Jmschrei/tangermeme, July 2025.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. DIME-FM : DIstilling Multimodal and Efficient Foundation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15533, 2023.

Amlan Talukder, Clayton Barham, Xiaoman Li, and Haiyan Hu. Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, 22(3):bbaa177, May 2021. ISSN 1477-4054. doi: 10.1093/bib/bbaa177.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K. Koo. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biology*, 26(1):203, July 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03674-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Raviteja Vemulapalli, Hadi Pouransari, Fartash Faghri, Sachin Mehta, Mehrdad Farajtabar, Mohammad Rastegari, and Oncel Tuzel. Knowledge Transfer from Vision Foundation Models for Efficient Training of Small Task-specific Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 49345–49367. PMLR, July 2024.

Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveen K. Kanithi, Marco AF Pimentel, Ronnie Rajan, and Shadab Khan. Genomic Foundationless Models: Pretraining Does Not Promise Performance, December 2024.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. Specialized Foundation Models Struggle to Beat Supervised Baselines. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, November 2024.

Chuanpeng Yang, Wang Lu, Yao Zhu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: Methods, evaluation, and application, 2024. URL `https://arxiv.org/abs/2407.01885`.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, Yonghong Tian, Jian Zhan, Jie Chen, and Yaoqi Zhou. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3, November 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1031.

Jessica Zhou, Kaeli Rizzo, Trevor Christensen, Ziqi Tang, and Peter K Koo. Uncertainty-aware genomic deep learning with knowledge distillation. *NPJ Artif. Intell.*, 2(1):3, January 2026.

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, October 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3547.

Tianxun Zhou and Keng-Hwee Chiam. Synthetic data generation method for data-free knowledge distillation in regression neural networks. *Expert Systems with Applications*, 227:120327, October 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.120327.

A APPENDIX

Task Configuration	NCR	APA	MRL	OFF	ON	ONF
RNA-based FM	RF	TE+EL	TE+EL	RF	MSM	TE+EL
Fine-tuning	True	True	True	True	True	True
KD Loss	EuCo	Fitnet	EuCo	EuCo	EuCo	EuCo
Loss Aggregation	Sum	Sum	Cvx	Sum	Cvx	Cvx

Table A.1: **Best OneT-KD configurations.** RNA-based FM: RNA-FM (RF), UTR-LM (TE+EL), UTR-LM (MRL), RNA-MSM (MSM). Fine-tuning: True or False. KD Loss: Eucosine (EuCo), Fitnet, Logits. Loss Aggregation: Sum or Convex (Cvx).

Hyperparameter	**Search Space**
FMs fine-tuning	
Batch size	$\{2, 4, 32, 128\}$
Learning rate	$\{10^{-3}, 10^{-4}, 10^{-5}\}$
Baseline CNN	
Embedding dimension	$\{32, 64, 128\}$
First layer channels (ch_{init})	$\{16, 32, 64, 128, 256, 512\}$
Kernel size (ks)	$\{3, 6, 9, 12\}$
Dropout (p_{drop})	$\mathcal{U}(0.0,\ 0.3)$
Learning rate	$\log \mathcal{U}(5 \times 10^{-4},\ 10^{-2})$
Batch size	$\{16, 32, 64\}$
α_{KD} in $\mathcal{L}_{convex}$	$\mathcal{U}(0.0,\ 1.0)$

Table A.2: **Hyperparameters Search Spaces.**

Model	FM	FT	Loss	Task Metric Agg	NCR MCC	NCR WF1	NCR BAS	NCR R2	APA PCC	APA MSE	MRL R2	MRL PCC	MRL MSE	OFF R2	OFF PCC	OFF MSE	ON R2	ON PCC	ONF R2	ONF PCC	ONF MSE
RNA-FM					96.38	96.63	96.65	86.73	93.62	1.24	82.42	90.88	13.67	35.18	65.41	5.68	67.13	83.45	2.84	33.12	63.48	5.71
RNA-MSM					90.73	91.36	91.42	88.21	93.92	1.22	79.57	89.95	15.88	31.71	62.57	5.99	66.55	82.54	2.89	36.39	62.75	5.43
UTR-LM (MRL)					86.73	97.64	87.73	88.81	94.25	1.16	82.05	90.94	13.96	47.11	69.12	4.64	73.10	85.73	2.32	44.42	67.51	4.74
UTR-LM (TE+EL)					82.64	83.25	83.88	88.53	94.13	11.90	82.94	91.35	13.27	49.76	71.00	4.40	73.26	85.92	2.31	45.50	68.14	4.65
HyenaDNA					82.39	80.76	83.49	81.98	90.56	1.78	67.01	82.77	25.66	38.88	62.45	5.36	61.81	78.63	3.30	31.58	56.31	5.84
ESM2					77.28	78.99	79.00	67.75	83.11	3.34	31.74	57.11	53.09	36.85	61.67	5.54	65.57	81.08	2.97	32.37	57.41	5.77
Baseline					84.15	84.43	85.27	87.74	93.69	1.27	74.80	86.66	19.60	50.26	70.90	4.36	74.14	86.12	2.23	45.73	67.75	4.63
OneT-KD	RNA-FM	F	EuCo	cvx	84.15	84.43	85.23	86.61	93.17	1.39	73.44	85.89	20.65	49.74	70.56	4.41	73.81	85.95	2.26	45.48	67.78	4.65
				sum	91.20	91.86	91.84	87.41	93.59	1.30	73.05	85.56	20.95	49.57	70.43	4.42	73.52	85.83	2.29	45.25	67.43	4.67
			fitnet	cvx	85.60	85.70	86.58	86.64	93.17	1.38	73.41	85.92	20.68	49.89	70.72	4.39	73.78	85.95	2.26	44.55	66.79	4.73
				sum	89.08	89.80	89.88	87.52	93.62	1.29	73.72	86.13	20.44	49.51	70.42	4.43	74.11	86.12	2.24	44.84	67.02	4.71
		T	EuCo	cvx	94.31	94.79	94.73	86.94	93.34	1.35	73.86	85.95	20.33	50.94	71.39	4.30	74.31	86.25	2.22	46.44	68.15	4.57
				sum	94.64	95.12	95.04	87.88	93.83	1.25	74.19	86.54	20.07	51.26	71.64	4.27	74.69	86.47	2.18	47.38	69.02	4.49
			fitnet	cvx	93.98	94.48	94.42	86.97	93.28	1.35	71.40	85.63	22.24	50.42	71.09	4.35	74.03	86.10	2.24	45.85	68.20	4.62
				sum	94.51	94.87	94.92	87.71	93.67	1.27	74.42	86.33	19.89	49.99	70.72	4.38	73.92	86.00	2.25	45.00	67.16	4.69
			logits	cvx	92.31	92.97	92.88	86.67	93.15	1.38	72.44	85.28	21.43	48.95	69.98	4.47	73.62	85.85	2.28	45.09	67.28	4.69
				sum	93.00	93.52	93.54	66.78	93.19	3.44	53.44	83.88	36.21	41.04	70.40	5.17	58.22	85.47	3.61	38.02	66.76	5.29
	RNA-MSM	F	EuCo	cvx	84.04	84.17	85.11	86.98	93.31	1.35	74.07	86.18	20.16	49.93	70.71	4.39	73.84	85.98	2.26	44.95	67.38	4.70
				sum	86.91	87.00	87.77	87.63	93.68	1.28	72.96	85.70	21.03	49.49	70.40	4.43	73.82	85.94	2.26	45.09	67.37	4.69
			fitnet	cvx	83.37	83.75	84.54	86.91	93.36	1.36	73.29	85.80	20.77	50.26	70.96	4.36	73.69	85.86	2.27	45.52	67.76	4.65
				sum	85.49	85.62	86.42	87.51	93.62	1.29	73.04	85.53	20.96	49.69	70.53	4.41	73.67	85.90	2.27	44.83	67.30	4.71
		T	EuCo	cvx	94.01	94.52	94.46	88.35	94.06	1.21	74.61	86.55	19.75	50.90	71.53	4.30	74.85	86.52	2.17	47.53	69.17	4.48
				sum	94.40	94.92	94.81	88.41	94.08	1.20	74.18	86.44	20.08	50.97	71.41	4.30	74.81	86.57	2.17	46.51	68.21	4.56
			fitnet	cvx	93.56	93.96	94.04	87.50	93.65	1.29	73.80	86.10	20.37	50.07	70.80	4.38	73.96	86.00	2.25	44.67	66.90	4.72
				sum	92.43	92.94	93.00	87.78	93.72	1.27	73.19	85.95	20.85	49.60	70.46	4.42	73.85	85.94	2.26	45.47	67.72	4.65
			logits	cvx	91.23	91.83	91.88	86.57	93.30	1.39	71.80	84.75	21.93	49.72	70.59	4.41	73.72	85.87	2.27	45.91	67.87	4.62
				sum	90.31	91.10	91.04	67.93	93.24	3.32	59.41	84.68	31.57	41.22	69.51	5.15	59.33	85.81	3.51	38.51	66.81	5.25
	UTR-LM (MRL)	F	EuCo	cvx	82.61	83.06	83.88	87.20	93.39	1.33	72.88	85.80	21.09	50.15	70.82	4.37	73.50	85.76	2.29	45.24	67.54	4.67
				sum	85.19	85.37	86.19	86.97	93.27	1.35	72.85	85.37	21.11	50.30	70.97	4.36	73.73	85.91	2.27	45.19	67.65	4.68
			fitnet	cvx	85.68	85.74	86.65	87.11	93.40	1.33	73.99	86.20	20.22	50.42	71.03	4.35	73.73	85.90	2.27	45.55	67.62	4.65
				sum	83.62	83.75	84.73	87.83	93.78	1.26	74.03	86.26	20.20	50.07	70.82	4.38	74.05	86.08	2.24	45.75	67.99	4.63
		T	EuCo	cvx	93.68	94.10	94.15	88.50	94.15	1.19	75.09	86.74	19.37	50.49	71.13	4.34	74.02	86.09	2.24	46.60	68.27	4.56
				sum	94.56	95.03	94.96	88.37	94.14	1.20	74.44	86.62	19.88	49.98	70.72	4.38	74.29	86.20	2.22	47.22	68.74	4.50
			fitnet	cvx	91.54	92.18	92.19	88.47	94.06	1.19	75.49	87.26	19.06	50.68	71.22	4.32	74.31	86.26	2.22	47.00	68.57	4.52
				sum	93.00	93.57	93.54	88.65	94.20	1.18	74.59	86.97	19.76	50.39	71.04	4.35	74.61	86.39	2.19	47.24	68.87	4.50
			logits	cvx	91.48	92.11	92.11	86.67	93.15	1.38	72.87	85.94	21.10	49.29	70.37	4.44	73.63	85.83	2.28	44.58	67.71	4.73
				sum	90.60	91.40	91.31	67.31	92.86	3.38	54.56	84.34	35.34	40.49	69.15	5.22	58.86	85.25	3.55	38.44	67.35	5.25
	UTR-LM (TE+EL)	F	EuCo	cvx	85.36	83.14	86.15	86.85	93.30	1.36	69.95	83.85	23.37	49.70	70.56	4.41	73.77	85.95	2.26	45.20	67.43	4.68
				sum	87.94	87.93	88.69	86.91	93.31	1.36	71.86	84.81	21.89	49.23	70.29	4.45	73.82	85.95	2.26	44.59	67.13	4.73
			fitnet	cvx	84.15	84.41	85.23	87.29	93.45	1.32	72.99	85.45	21.01	50.02	70.84	4.38	73.49	85.79	2.29	45.01	67.45	4.69
				sum	81.11	81.69	82.42	87.39	93.58	1.31	73.38	86.04	20.70	49.61	70.53	4.42	74.01	86.05	2.24	44.97	67.10	4.70
		T	EuCo	cvx	86.14	83.76	86.88	88.78	94.27	1.16	77.24	88.09	17.70	50.87	71.33	4.31	74.33	86.25	2.22	47.78	69.15	4.46
				sum	94.15	94.72	94.58	88.77	94.23	1.16	76.50	87.62	18.28	50.37	71.18	4.35	74.50	86.32	2.20	47.00	68.58	4.52
			fitnet	cvx	87.96	88.85	88.84	88.22	93.96	1.22	76.54	87.74	18.24	50.48	71.06	4.34	74.84	86.55	2.17	46.99	68.57	4.52
				sum	92.44	93.11	93.00	88.80	94.27	1.16	74.18	86.20	20.08	50.80	71.37	4.31	74.66	86.44	2.19	46.87	68.53	4.53
			logits	cvx	91.37	92.10	92.00	86.66	93.20	1.38	72.55	85.51	21.35	49.52	70.46	4.43	73.58	85.91	2.28	45.47	67.60	4.65
				sum	83.54	83.92	84.57	68.13	92.32	3.30	54.70	84.74	35.23	40.59	69.31	5.21	58.66	85.42	3.57	38.69	66.98	5.23
EnsAvg-RNA					94.46	94.90	94.88	90.21	94.98	1.01	85.10	92.31	11.59	51.84	72.10	4.22	75.31	86.84	2.13	48.91	69.95	4.35
EnsKD-RNA	All RNA	T	EuCo	sum	94.77	95.22	95.15	88.68	94.25	1.17	74.53	86.95	19.81	51.12	71.54	4.28	74.81	86.55	2.17	47.79	69.20	4.46
EnsAvg-MM					94.43	94.88	94.84	84.84	92.50	1.56	74.55	88.49	19.78	48.68	69.89	4.49	71.94	84.87	2.42	43.43	66.15	4.82
EnsKD-MM	MM	T	EuCo	sum	93.31	93.99	93.77	88.46	94.09	1.19	73.00	85.75	20.99	50.80	71.28	4.31	74.36	86.27	2.21	46.45	68.32	4.57

Table A.3: **Full ablation results**. Per task model performance across the different experimental settings. This covers the four RNA-based FMs, the DNA and protein FMs, the supervised baseline (Baseline), as well as the different KD settings - OneT-KD, EnsAvg-RNA, EnsKD-RNA, EnsAvg-MM, and EnsKD-MM. In the OneT-KD case, results for the different teachers, finetuned (FT) or not (T=True, F=False), as well as different KD losses and loss aggregation functions are included. The color gradients highlight the top-3 results per column.

Table A.4: **Best Hyperparameters**. Optimal hyperparameter configurations obtained from Ray Tune for the supervised baseline and the distilled student models, along with the grid search results for the foundation model fine-tuning. Per-task CSV files are provided with the code repository.

Table A.5: **Motif functionality analysis**. A comprehensive list of the GO Biological processes, KEGG, and Reactome terms that were significantly enriched in the overrepresentation analysis, including the adjusted p-values and enrichment scores are available on the code repository.