Investigating Data Pruning for Pretraining Biological Foundation Models at Scale

Yifan Wu[1,2,*], **Jiyue Jiang**[1,*,†], **Xichen Ye**[2], **Yiqi Wang**[2], **Chang Zhou**[1], **Yitao Xu**[1], **Jiayang Chen**[1], **He Hu**[3], **Weizhong Zhang**[2,†], **Cheng Jin**[2], **Jiao Yuan**[4,5,†], **Yu Li**[1,†]

[1]The Chinese University of Hong Kong
[2]Fudan University
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[4]Guangzhou National Laboratory
[5]Guangzhou Medical University
{victorwu, jiangjy}@link.cuhk.edu.hk, weizhongzhang@fudan.edu.cn, yuan_jiao@gzlab.ac.cn, liyu@cse.cuhk.edu.hk

Abstract

Biological foundation models (BioFMs), pretrained on large-scale biological sequences, have recently shown strong potential in providing meaningful representations for diverse downstream bioinformatics tasks. However, such models often rely on millions to billions of training sequences and billions of parameters, resulting in prohibitive computational costs and significant barriers to reproducibility and accessibility—particularly for academic labs. To address these challenges, we investigate the feasibility of data pruning for BioFM pretraining and propose a post-hoc influence-guided data pruning framework tailored to biological domains. Our approach first introduces a subset-based self-influence formulation that enables efficient estimation of sample importance at low computational cost. Built upon this, we propose two simple yet effective selection strategies: Top-k Influence (Top I) and Coverage-Centric Influence (CCI). Then, we empirically validate our method on two representative BioFMs: RNA-FM and ESM-C. For RNA, our framework consistently outperforms random selection baselines under an extreme pruning rate of over 99%, which displays our framework's effectiveness. Furthermore, we demonstrate the generalizability of our framework on protein-related tasks using ESM-C. Specifically, our coreset even outperforms random 10x subsets in both RNA and protein settings, revealing substantial redundancy in biological sequence datasets. These findings underscore the potential of influence-guided data pruning to substantially reduce the computational cost of BioFM pretraining, paving the way for more efficient, accessible, and sustainable biological AI research.

Code — https://github.com/victor-yifanwu/bio-coreset

1 Introduction

Recent advances in biological foundation models (BioFMs) have enabled remarkable progress in tasks such as structure prediction, functional annotation, and molecular interaction modeling across RNA/DNA and protein sequences (Chen et al. 2022; Brixi et al. 2025; Hayes et al. 2025; Shen et al. 2024; Team et al. 2025). Despite their success, these models typically rely on extremely large-scale pretraining data, which demand substantial computational, environmental, and reproducibility costs. For example, RNA-FM is trained on over 23 million RNA sequences (Chen et al. 2022), while ESM has scaled up to 2.78 billion protein sequences in its latest versions (Hayes et al. 2025), making it practically infeasible for most research groups to reproduce the full training process.

To promote open and sustainable development in BioFMs, we focus on investigating the potential of data pruning (Phillips 2017) as a means to reduce substantial computational overhead while maintaining competitive performance. Specifically, we explore whether a carefully selected coreset can be used to retrain from scratch with significantly fewer examples. To the best of our knowledge, this direction has received little attention in the context of BioFMs pretraining. While coreset selection has been studied in CV/NLP (Moser et al. 2025; Diddee and Ippolito 2025), most existing approaches fall into two main categories: (i) those that rely on training dynamics (Pleiss et al. 2020; He et al. 2024); (ii) those based on local density measures (Yang et al. 2024; Zhang et al. 2025). However, both types face fundamental limitations in the biological setting. First, the pretraining cost is prohibitively high, and most BioFMs do not publicly release training details, rendering training-dynamics-based methods inapplicable. Second, the millions-to-billions scale of biological sequences poses significant scalability barriers for methods that depend on pairwise similarity computations.

To address these limitations, we propose an influence-guided coreset selection framework that operates in a post-hoc manner, without requiring access to the full training process. Specifically, our framework consists of two main stages: (i) estimating influence scores for individual training examples; (ii) selecting subsets based on these scores with tailored selection strategies. First, grounded in the classical influence function framework (Koh and Liang 2017), we reformulate a scalable subset-based self-influence function that estimates the impact of each training example, replacing the need to compute Hessians over the entire training data. To make this approximation theoretically sound, we

[*]These authors contributed equally to this work.
[†]Corresponding authors.

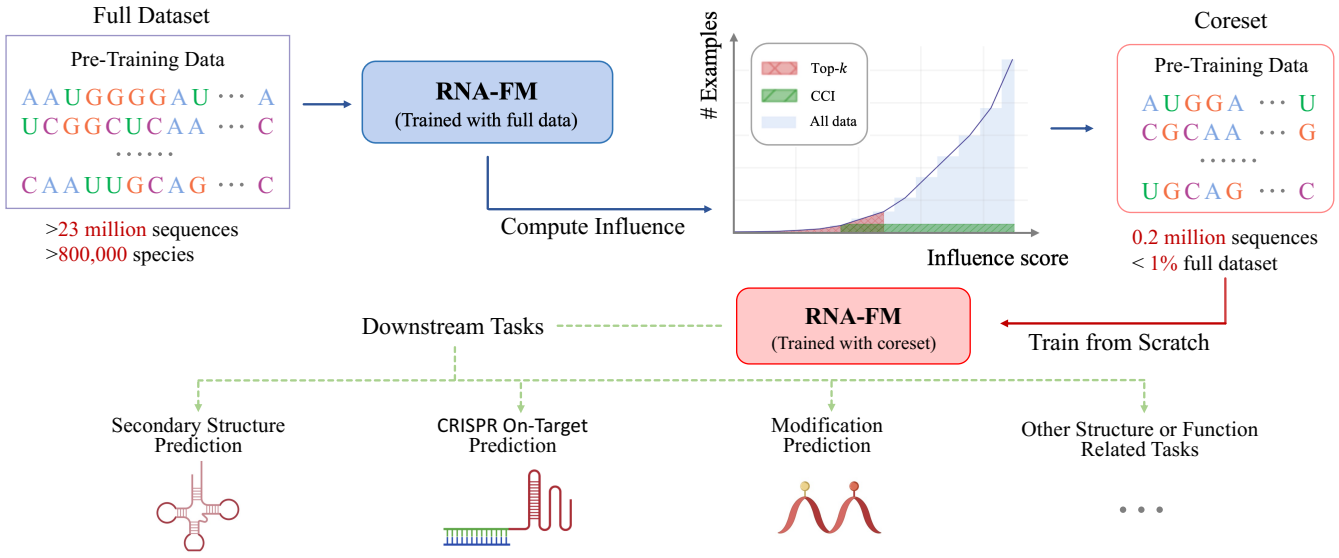

Figure 1: An overview of our proposed influence-guided coreset selection and evaluation pipeline for large-scale RNA sequence pretraining.

introduce a key assumption—the subset-based ERM condition—which requires the model to be sufficiently trained on a small randomly sampled subset. When this condition is satisfied, the curvature around the subset can serve as a faithful surrogate for the full training curvature. To further reduce the computational cost, we adopt a diagonal empirical Fisher matrix as a tractable curvature approximation, enabling scalable influence estimation even at the scale of biological foundation models. Next, we introduce two influence-guided selection strategies designed to serve different objectives: (1) Top-k Influence-guided Selection (Top I); (2) Coverage-centric Influence-guided Selection (CCI). These two strategies allow us to explore how influential or diverse training examples contribute to representation learning. To evaluate their effectiveness, we pretrain BioFMs on both RNA and protein sequences using only 0.2 million selected examples in each case and assess their performance across a comprehensive suite of downstream tasks. An overview of our influence-guided coreset framework and evaluation pipeline is illustrated in Figure 1, using RNA-FM as a representative example. Our contributions are listed as follows:

- We propose a post-hoc influence-guided data pruning framework tailored for biological foundation models, eliminating the need for full training access.
- We provide a theoretical derivation of our reformulated influence function, which supports efficient approximation through curvature over randomly sampled subsets.
- We introduce two influence-guided coreset selection strategies: Top I and CCI, both of which help us to understand the representation ability of corresponding coresets.
- We first conduct extensive experiments on RNA-FM, demonstrating that our data pruning framework achieves competitive performance across multiple downstream tasks, and then validate its generalizability on ESM-C.

2 Related Work

2.1 Data Pruning

Data pruning, also known as coreset selection, aims to identify a small yet representative subset of a large training corpus such that training on this subset yields performance comparable to using the full dataset (Phillips 2017; Moser et al. 2025). One common approach assigns importance scores to individual training examples based on training dynamics—for example, EL2N (Coleman et al. 2020), AUM (Pleiss et al. 2020), and Dynamic Uncertainty (He et al. 2024). These methods quantify informativeness from model prediction behaviors over the course of training and are efficient when the full training process is accessible. However, they require tracking model outputs across multiple epochs, which is often infeasible or prohibitively expensive in the context of BioFMs. Another line of work focuses on local data density (Yang et al. 2024), where representativeness is measured by a sample's proximity to its neighbors in the feature space. Yet such approaches typically rely on storing high-dimensional embeddings and performing pairwise similarity computations, which becomes computationally impractical for large-scale datasets. An alternative line of research is grounded in influence functions (Koh and Liang 2017), which will be further discussed in Section 3.1. While theoretically grounded, influence functions require computing gradients and inverse Hessian, which becomes computationally intensive in large-scale settings.

2.2 Biological Sequence Representation

Learning effective representations of biological sequences—RNA, DNA, and proteins—is fundamental to a

wide range of downstream tasks, including structure prediction, function annotation, and biomolecular interaction modeling (Shen et al. 2024; Wang et al. 2025). Due to the intrinsic complexity of biological macromolecules, such as RNA's hierarchical structure and protein folding patterns, traditional approaches often relied on hand-crafted features or shallow learning techniques specific to individual tasks. Recent advances in biological foundation models (BioFMs) (Chen et al. 2022; Hayes et al. 2025; Brixi et al. 2025) have demonstrated the effectiveness of large-scale self-supervised learning, particularly masked language modeling, in capturing intricate biological semantics directly from raw sequences. For RNA, models, such as RNA-FM (Chen et al. 2022), have achieved strong transfer performance across diverse tasks, including RNA type classification (Amin, McGrath, and Chen 2019), CRISPR-Cas efficiency prediction (Chuai et al. 2018), and RNA-binding protein (RBP) interaction prediction (Xu et al. 2023). On the protein side, ESM (Hayes et al. 2025) has shown promising results on protein function prediction (Sarkisyan et al. 2016), protein structure prediction (Klausen et al. 2019), and interaction prediction (Guo et al. 2008). Despite their success, BioFMs suffer from high training costs and limited reproducibility, which further motivates the need for post-hoc data-efficient approaches, such as data pruning, that can identify informative training subsets without full retraining access.

3 Methods

3.1 Preliminaries: Influence Function

Influence functions (IF) aim to understand the effect of individual training points on a model's predictions, which can be instantiated as the task of estimating how the model's output would change if a particular training point were removed (Hampel 1974).

Let $D_{\text{tr}} = \{z_n = (x_n, y_n)\}_{n=1}^N$ be an i.i.d. training dataset. Empirical risk minimization (ERM) solves:

$$\theta^\star := \arg\min_\theta \frac{1}{N}\sum_{n=1}^N \ell(z_n, \theta),$$

where ℓ denotes a per-sample loss function and $\theta^\star$ is the resulting minimizer. Next consider a validation set $D_{\text{val}} = \{z_m = (x_m, y_m)\}_{m=1}^M$. The influence of a training point $z_{\text{tr}} \in D_{\text{tr}}$ on a specific validation example $z_{\text{val}} \in D_{\text{val}}$ can be expressed as the excess loss: $\ell(z_{\text{val}}, \theta^\star_{z_{\text{tr}}}) - \ell(z_{\text{val}}, \theta^\star)$, where $\theta^\star_{z_{\text{tr}}}$ is the solution to a perturbed ERM objective in which z_{tr} is upweighted by a small amount ϵ:

$$\theta^\star_{z_{\text{tr}}} := \arg\min_\theta \frac{1}{N}\sum_{n=1}^N \ell(z_n, \theta) + \epsilon\ell(z_{\text{tr}}, \theta).$$

In particular, setting $\epsilon = -\frac{1}{N}$ corresponds to the removal of z_{tr} from the training set.

Following Koh and Liang (2017), this excess loss can be approximated via a two-step procedure.

Step 1: Parameter change. The shift in parameters due to perturbing z_{tr} can be approximated by a Newton step:

$$\theta^\star_{z_{\text{tr}}} - \theta^\star \approx -\epsilon H_{\text{tr}}^{-1} g_{z_{\text{tr}}},$$

where $g_{z_{\text{tr}}} = \nabla_{\theta^\star}\ell(z_{\text{tr}}, \theta^\star)$ denotes the gradient of the loss w.r.t. z_{tr}, and $H_{\text{tr}} = \frac{1}{N}\sum_{n=1}^N \nabla^2_{\theta^\star}\ell(z_n, \theta^\star)$ is the Hessian of the empirical loss over D_{tr}, which captures the local curvature of the empirical risk around the training optimum (Van der Vaart 2000).

Step 2: Loss change. The change in the validation loss of the sample z_{val} can then be estimated via first-order Taylor expansion:

$$\ell(z_{\text{val}}, \theta^\star_{z_{\text{tr}}}) - \ell(z_{\text{val}}, \theta^\star) \approx g_{z_{\text{val}}}^\top(\theta^\star_{z_{\text{tr}}} - \theta^\star),$$

where $g_{z_{\text{val}}} = \nabla_{\theta^\star}\ell(z_{\text{val}}, \theta^\star)$.

Final form. Combining the two steps yields the influence function of z_{tr} to z_{val} on $\theta^\star$:

$$\mathcal{I}(z_{\text{tr}}; z_{\text{val}}) := g_{z_{\text{val}}}^\top H_{\theta^\star}^{-1} g_{z_{\text{tr}}}. \tag{1}$$

In addition, the influence function of z_{tr} on the entire validation set D_{val} can be extended as follows:

$$\mathcal{I}(z_{\text{tr}}; D_{\text{val}}) := \frac{1}{M}\sum_{m=1}^M g_{z_m}^\top H_{\text{tr}}^{-1} g_{z_{\text{tr}}}, \tag{2}$$

where $z_m \in D_{\text{val}}$ represents a validation sample.

By replacing D_{val} with the training set D_{tr} in Equation (2), we obtain the self-influence function (Koh and Liang 2017) of a training point:

$$\mathcal{I}(z_{\text{tr}}; D_{\text{tr}}) := \frac{1}{N}\sum_{n=1}^N g_{z_n}^\top H_{\text{tr}}^{-1} g_{z_{\text{tr}}}, \tag{3}$$

where $z_n \in D_{\text{tr}}$. This allows us to quantify the effect of each training point in self-supervised settings. However, computing influence scores over large-scale training sets remains prohibitively expensive in both memory and computation. To address this, we introduce scalable approximations in the following sections.

3.2 Subset-Based Scalable Influence Estimation

Direct computation of the standard self-influence function, as defined in Eq. 3, requires accessing the full training Hessian H_{tr} and computing its inverse. For BioFMs like RNA-FM, which contain billions of parameters, this becomes infeasible in both memory and runtime. Motivated by recent advances in curvature-based influence reformulations (Ye et al. 2025), we investigate whether influence can be approximated using curvature estimated over a small training subset. To this end, we propose a two-step strategy: (i) reformulate the influence function based on the training subset; (ii) apply a lightweight inverse Hessian approximation to further reduce computational overhead.

Reformulating Self-Influence via the Training Subset The classical self-influence function (Koh and Liang 2017) assumes the model is trained to minimize empirical risk over the full training set, thereby allowing the influence of a single example to be approximated via the curvature of the training loss, i.e., H_{tr}. To enable scalable estimation, we propose to replace H_{tr} with H_{sub} computed on a small random sampled training subset. To do so, we first let $D_{\text{sub}} = \{z_m\}_{m=1}^M \subset D_{\text{tr}}$ be a randomly sampled subset of

the training set. The model parameters $\tilde{\theta}$ then can be obtained by minimizing the empirical risk over D_{sub} similar to Koh and Liang (2017):

$$\tilde{\theta} := \arg\min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \ell(z_m, \theta). \tag{4}$$

Next, as described in Section 3.1, we also decompose self-influence estimation into two steps: parameter change and loss change.

Step 1: Parameter change. Different from $\theta^{\star}_{z_{\text{tr}}}$ as described in Section 3.1, we consider $\tilde{\theta}_{z_{\text{tr}}}$ defined as:

$$\tilde{\theta}_{z_{\text{tr}}} := \arg\min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \ell(z_m, \theta)) + \epsilon \ell(z_{\text{tr}}, \theta).$$

Therefore, the parameter change $\tilde{\theta}_{z_{\text{tr}}} - \tilde{\theta}$ can be approximated using a Newton step:

$$\tilde{\theta}_{z_{\text{tr}}} - \tilde{\theta} \approx -\epsilon \tilde{H}_{\text{sub}}^{-1} \tilde{g}_{z_{\text{tr}}}, \tag{5}$$

where $\tilde{g}_{z_{\text{tr}}} = \nabla_{\tilde{\theta}} \ell(z_{\text{tr}}, \tilde{\theta})$ is the gradient of the loss w.r.t. $\tilde{\theta}$ for training poit z_{tr} and $\tilde{H}_{\text{sub}} = \frac{1}{M} \sum_{m=1}^{M} \nabla^2_{\tilde{\theta}} \ell(z_m, \tilde{\theta})$ is the Hessian matrix computed over D_{sub}.

Step 2: Loss change. Previous studies (Kim, Kim, and Yang 2023) have shown that second-order approximations yield more accurate estimates of loss changes. Therefore, we adopt a second-order Taylor expansion to estimate the loss change:

$$\begin{aligned} &\ell(z_{\text{sub}}, \tilde{\theta}_{z_{\text{tr}}}) - \ell(z_{\text{sub}}, \tilde{\theta}) \\ \approx\ &\tilde{g}_{z_{\text{sub}}}^{\top} \Delta\theta + \frac{1}{2} \Delta\theta^{\top} \nabla^2_{\tilde{\theta}} \ell(z_{\text{sub}}, \tilde{\theta}) \Delta\theta, \end{aligned} \tag{6}$$

where $\Delta\theta = \tilde{\theta}_{z_{\text{tr}}} - \tilde{\theta}$ and $\tilde{g}_{z_{\text{sub}}} = \nabla_{\theta} \ell(z_{\text{sub}}, \tilde{\theta})$.

Final form. Combining Eq. 5 and Eq. 6, we can have self-influence function of z_{tr} to z_{sub} on $\tilde{\theta}$:

$$\begin{aligned} \mathcal{I}(z_{\text{tr}}, z_{\text{sub}}) := &\tilde{g}_{z_{\text{sub}}}^{\top} \tilde{H}_{\text{sub}}^{-1} \tilde{g}_{z_{\text{tr}}} \\ &+ \frac{1}{2} \epsilon \tilde{g}_{z_{\text{tr}}}^{\top} \tilde{H}_{\text{sub}}^{-1} \nabla^2_{\tilde{\theta}} \ell(z_{\text{sub}}, \tilde{\theta}) \tilde{H}_{\text{sub}}^{-1} \tilde{g}_{z_{\text{tr}}}. \end{aligned} \tag{7}$$

With Eq. 4, i.e., $\tilde{g}_{\text{sub}} \to 0$, we will directly drop $\tilde{g}_{z_{\text{sub}}}^{\top} \tilde{H}_{\text{sub}}^{-1} \tilde{g}_{z_{\text{tr}}}$ later. By extending Eq. 7, we can measure the self-influence of z_{tr} on D_{sub}:

$$\mathcal{I}(z_{\text{tr}}, D_{\text{sub}}) := \tilde{g}_{z_{\text{tr}}}^{\top} \tilde{H}_{\text{sub}}^{-1} \tilde{g}_{z_{\text{tr}}}. \tag{8}$$

Given the uniform training objective and the flat loss landscape observed in large models (Chen et al. 2025), the curvature over a random subset D_{sub} is expected to approximate that of the full training set D_{tr}. Therefore, the self-influence of z_{tr} on D_{tr} can be approximated by $\mathcal{I}(z_{\text{tr}}, D_{\text{sub}})$.

Efficient Approximation for Inverse Hessian Despite the reformulation in Eq. 8, computing the full Hessian inverse H_{sub}^{-1} remains computationally intractable. The total complexity amounts to $O(M \cdot d^2 + d^3)$, where M is the number of subset examples and d is the number of model parameters. This cost is prohibitive for large-scale BioFMs with billions of parameters.

To address this issue, we note that BioFMs are typically trained with a negative log-likelihood loss and, under Eq 4, the model is well-trained on the subset, allowing us to approximate the Hessian H_{sub} using the empirical Fisher information matrix (Pascanu and Bengio 2014). Specifically, the empirical Fisher matrix over the training subset is given by:

$$\tilde{F}_{\text{sub}} := \frac{1}{M} \sum_{m=1}^{M} \tilde{g}_{z_m} \tilde{g}_{z_m}^{\top}, \tag{9}$$

where $\tilde{g}_{z_m} := \nabla_{\tilde{\theta}} \ell(z_m, \tilde{\theta})$ denotes the gradient of the loss w.r.t. the pretrained model parameters $\tilde{\theta}$ on the subset sample $z_m \in D_{\text{sub}}$.

To further reduce computational cost, we follow the practice commonly adopted in conjugate gradient methods (Roux, Manzagol, and Bengio 2007; Schaul, Zhang, and LeCun 2013; Martens and Grosse 2015) and adaptive optimizers such as Adam (Kingma and Ba 2015), where the curvature matrix is approximated by its diagonal. In this spirit, we apply a diagonal approximation to the empirical Fisher matrix, which yields:

$$\text{diag}(\tilde{F}_{\text{val}}) := \frac{1}{M} \sum_{m=1}^{M} \tilde{g}_{z_m} \odot \tilde{g}_{z_m}, \tag{10}$$

where $\odot$ denotes the element-wise (Hadamard) product. Therefore, the inverse Hessian can be approximated as:

$$\tilde{H}_{\text{val}}^{-1} \approx \text{diag}(\tilde{F}_{\text{val}})^{-1}. \tag{11}$$

Substituting this into Eq. 8, we obtain the final scalable approximation for subset-based self-influence function:

$$\mathcal{I}(z_{\text{tr}}, D_{\text{sub}}) := \tilde{g}_{z_{\text{tr}}}^{\top} \text{diag}(\tilde{F}_{\text{sub}})^{-1} \tilde{g}_{z_{\text{tr}}}. \tag{12}$$

This approximation enables influence estimation with linear complexity, reducing the overall computational cost from $O(M \cdot d^2 + d^3)$ to $O(M \cdot d)$—making it practical for large-scale BioFMs with billions of parameters.

Remark. The effectiveness of both the subset-based influence approximation and the Fisher-based curvature estimation hinges on Eq. 4, which assumes that the model is well-trained on the selected subset. In practice, we find that a light-weight fine-tuning (e.g., one epoch) on the subset is sufficient to satisfy this condition. The associated cost is negligible compared to the overall influence estimation pipeline, making the approach practical for large-scale BioFMs.

3.3 Influence-guided Coreset Selection Strategy

Building on the theoretical properties of influence functions (Koh and Liang 2017), we propose two selection strategies: Top-k Influence-guided Selection and Coverage-centric Influence-guided Selection, tailored for coreset construction.

Top-k Influence-guided Selection. Since the influence score of a training example quantifies its estimated contribution to the model's performance, selecting the top-k examples with the highest influence naturally prioritizes those with the greatest potential to affect generalization. This simple yet principled approach aligns with the coreset selection goal of retaining the most informative points.

Coverage-centric Influence-guided Selection. Recent findings (Sorscher et al. 2022; Xia et al. 2023) have revealed that the utility of different examples depends on the available data regime. Specifically, when only a small amount of data is retained, it is often more effective to preserve the *easiest* examples, as they convey coarse-grained information about the target function and help avoid overfitting. In contrast, hard examples typically provide fine-grained information, which becomes useful only when the model has already captured the basics of the distribution. Under extreme pruning, focusing solely on the hardest or most influential examples may hinder learning, as outliers or rare cases may dominate the subset while the underlying structure of the data remains underrepresented (Swayamdipta et al. 2020). Motivated by this, we apply stratified sampling over the influence score distribution, which ensures both easy and hard examples remain under extreme pruning.

Comparison. Top-k influence-guided selection emphasizes informativeness and parameter sensitivity, whereas Coverage-centric Influence-guided selection focuses on representational diversity and robustness under high pruning rates. To further explore these two strategies, we empirically evaluate both strategies in Section 4, and discuss when each approach may be preferable.

4 Experiments

To validate the effectiveness and generalizability of our post-hoc influence-guided data pruning framework, we conduct experiments on both RNA and protein foundation models, namely RNA-FM (Chen et al. 2022) and ESM-C (Hayes et al. 2025). Considering the prohibitive cost of pretraining BioFMs[1], we adopt an extreme data pruning evaluation setting, where only 0.2 million sequences are retained and used for pretraining. For RNA-FM, we conduct data pruning over the entire 23M-sequence training data, i.e., over 99% data pruning. In contrast, given the inaccessibility of the full 2.78-billion protein sequence used for ESM-C pretraining, we instead collect around 4.5 million protein sequences from UniRef50 (Suzek et al. 2007) and conduct data pruning over this. After data pruning via different selection strategies, we pretrain BioFMs on 0.2 million sequences from scratch for 10 epochs and evaluate them across a range of downstream tasks.

4.1 Selection Strategies and Baselines

We consider the following selection strategies and baselines in our experiments:

- **RNA-FM / ESM-C**: The original foundation model trained on the full dataset.
- **Raw**: Untrained model.
- **Random**: Uniform random sampling of 0.2M sequences (matching our pruning budget) or 2M sequences.
- **Top I**: Applies Top-k influence-guided selection with our subset-based self-influence.
- **CCI**: Applies coverage-centric influence-guided selection with our subset-based self-influence.

For Top I and CCI, we consider two variants: the default version performs a lightweight fine-tuning step on a randomly selected 0.2-million training subset to adapt influence estimation, whereas the (w/o ft) variant uses scores computed from the initial pretrained model without any adaptation.

4.2 Downstream Tasks and Evaluation Metrics

To comprehensively evaluate model performance, we assess RNA-FM and ESM-C variants pretrained on different coresets, as well as full-data and raw baselines raw and full-data baselines, across a diverse suite of downstream RNA and protein understanding tasks.

RNA-FM The downstream tasks (Chen et al. 2022; Ren et al. 2024) can be categorized as follows:

- **Function Prediction**: (1) RNA Type Classification (TypeCls) (Amin, McGrath, and Chen 2019), evaluated using accuracy and F1 score; (2) RNA Modification Prediction (Modif) (Duan, Wang, and Jia 2019), evaluated using AUC.
- **Engineering Prediction**: CRISPR On-Target Prediction (CRI-On) (Chuai et al. 2018), evaluated using SC and MSE.
- **Structure Prediction**: (1) Secondary Structure Prediction on bpRNA (Danaee et al. 2018), evaluated using precision, recall, F1 score, and MCC; (2) Distance Map Prediction (Chen et al. 2022), evaluated using R^2, Spearman correlation (SC), MAE, and MSE; (3) Contact Map Prediction (Chen et al. 2022), evaluated using Top-L precision.
- **Interaction Prediction**: RBP-RNA Interaction prediction (Chen et al. 2022), evaluated using accuracy, AUC, and AUPR.

ESM The downstream tasks (Xu et al. 2022) can be categorized as follows:

- **Localization Prediction**: Binary Localization Prediction (Bin) (Almagro Armenteros et al. 2017), evaluated using accuracy.
- **Structure Prediction**: Secondary Structure Prediction (SS) (Klausen et al. 2019), evaluated using accuracy.
- **Interaction Prediction**: PPI Affinity Prediction(Aff) (Moal and Fernández-Recio 2012), evaluated using MAE and RMSE.

5 Results and Analysis

In this section, we first conduct a comprehensive evaluation of our influence-guided data pruning framework on RNA-FM, covering a broad range of RNA-specific tasks. To assess its generalizability, we further apply the same framework to ESM-C, a protein foundation model, demonstrating its robustness across distinct biomolecular modalities.

[1]RNA-FM was trained on 23 million sequences using 8 A100 GPUs over 30 days (Chen et al. 2022).

Methods	Data Size	TypeCls		Modif	CRI-On	
		ACC(%)	F1(%)	AUC(%)	SC(%)	MSE ↓
RNA-FM	23M	91.93	91.87	94.98	31.87	.0118
Raw	0M	79.46	78.96	90.71	22.48	.0261
Random	2M	82.21	82.01	92.82	26.72	.0158
Random	0.2M	82.15	81.97	91.86	26.67	.0161
Top I (w/o ft)	0.2M	81.07	81.21	92.94	<u>28.60</u>	.0151
CCI (w/o ft)	0.2M	80.60	80.37	<u>93.31</u>	26.96	.0150
Top I	0.2M	<u>82.51</u>	<u>82.53</u>	93.20	27.08	<u>.0149</u>
CCI	0.2M	**82.88**	**83.12**	**93.86**	**32.90**	**.0135**

Table 1: Performance of different coresets across three function and engineering prediction tasks (RNA Type Classification, RNA Modification Prediction, and CRISPR On-Target Prediction). **Bold** denotes the best results and <u>underline</u> denotes the second-best results.

5.1 Results on RNA-FM

Overall Performance across RNA Tasks We mainly categorize RNA downstream tasks into two groups: Function and Engineering Prediction and Structure and Interaction Prediction.

• **Function and Engineering Prediction** We first examine RNA Type Classification (TypeCls), Modification Prediction (Modif), and CRISPR On-Target Prediction (CRI-On). The complete results are reported in Table 1. Across all three tasks, both Top I and CCI consistently outperform the Random baseline. Although the performance margins of TypeCls over Random may appear modest, achieving consistent improvements across all tasks still demonstrates the effectiveness of our influence-guided approach and highlights the value of exploring informed pruning for BioFMs. Notably, CCI consistently achieves the best performance across all three tasks. Its superior results suggest that incorporating coverage and diversity patterns is more beneficial for the model to capture functional patterns of RNA sequences during self-supervised pretraining. Furthermore, on CRISPR On-Target Prediction, CCI even surpasses the full RNA-FM model trained on all 23 million sequences, which demonstrates that, in certain scenarios, high-quality, task-relevant information can be effectively preserved within a drastically reduced training subset.

• **Structure and Interaction Prediction** We then examine Secondary Structure Prediction on bpRNA, Distance Map Prediction, Contact Map Prediction, and RBP-RNA Interaction Prediction. The complete results are reported in Table 2. Both Top I and CCI again surpass the Random baseline across nearly all metrics, reinforcing the effectiveness of our influence-guided data pruning approach. Interestingly, Top I demonstrates notable advantages in structure- and interaction-related tasks, both of which are highly dependent on the underlying RNA structural properties. In particular, in Contact Map Prediction (Table 2c), Top I even surpasses RNA-FM on both Top-1.0L and Top-0.5L precision metrics. This suggests that, within RNA datasets, examples with higher self-influence scores tend to encode richer structural information, which can be effectively leveraged by models during self-supervised pretraining. Moreover, across all four tasks in this category, the performance gap between Top I and full-data RNA-FM remains remarkably small. This indicates that, for structure- and interaction-related tasks, a compact subset consisting of the most self-influential samples can effectively replace the full 23-million-sequence dataset for RNA-FM pretraining.

Data Redundancy in RNA Training Data From the results reported in Table 1, we observe that both Top I and CCI consistently outperform the Random 2M baseline across all function and engineering prediction tasks, while using only 10% of the data volume (0.2M vs. 2M sequences). This performance gain, achieved under a significantly smaller data budget, provides strong evidence that the RNA training data contains considerable redundancy. Such empirical evidence highlights the potential and necessity of exploring data pruning or coreset selection techniques tailored to RNA pretraining, especially under large-scale pretraining scenarios.

Ablation: Necessity of Adaptation for Subset-based Influence To validate the necessity of fine-tuning on the subset prior to influence estimation, we compare our influence-guided selection strategies with and without adaptation (denoted as w/o ft) in both Table 1 and Table 2. In all cases, the adapted variants (Top I and CCI) consistently outperform their non-adapted counterparts, demonstrating the importance of Eq. 4 before computing influence. These results confirm that the lightweight adaptation step is critical for reducing estimation error in both influence scores and curvature approximations.

5.2 Results on ESM

To assess the generalizability of our influence-guided data pruning framework, we apply it to the protein foundation model ESM-C. We evaluate model performance on three representative downstream tasks: Binary Localization Prediction (Bin), Secondary Structure Prediction (SS), and Protein–Protein Interaction Affinity Prediction (Aff). The complete results are displayed in Table 3. Our influence-guided data pruning strategies—Top I and CCI—consistently outperform both Random baselines (with 0.2M and 2M samples) across all three tasks, thereby verifying the effectiveness of our pruning framework in the protein domain. More-

Methods	Data Size	Pre(%)	Rec(%)	F1(%)	MCC(%)
RNA-FM	23M	66.14	62.24	62.20	63.01
Random	0.2M	59.75	55.59	55.60	56.49
Top I (w/o ft)	0.2M	59.74	58.22	56.95	57.76
CCI (w/o ft)	0.2M	59.30	57.20	56.10	57.00
Top I	0.2M	59.76	**58.27**	**57.05**	**57.85**
CCI	0.2M	**60.29**	56.33	56.36	57.14

(a) Secondary Structure Prediction on bpRNA.

Methods	Data Size	R^2(%)	SC(%)	MAE ↓	MSE ↓
RNA-FM	23M	83.26	89.21	.5665	.6650
Random	0.2M	76.71	84.90	.7176	1.037
Top I (w/o ft)	0.2M	75.91	84.13	.7284	1.057
CCI (w/o ft)	0.2M	76.80	84.95	.7254	1.045
Top I	0.2M	**79.25**	**86.47**	**.6745**	**.9215**
CCI	0.2M	77.98	85.59	.6937	.9861

(b) RNA Distance Map Prediction.

Methods	Data Size	Long-Range Top Precision (%)			
		L:1.0L	L:0.5L	L:0.2L	L:0.1L
RNA-FM	23M	93.93	98.28	99.62	99.86
Random	0.2M	94.18	98.20	99.28	99.31
Top I (w/o ft)	0.2M	93.94	98.05	99.06	98.99
CCI (w/o ft)	0.2M	93.86	98.22	99.32	**99.46**
Top I	0.2M	**94.36**	**98.41**	**99.39**	99.39
CCI	0.2M	94.20	98.26	99.14	99.21

(c) RNA Contact Map Prediction.

Methods	Data Size	ACC(%)	AUPR(%)	AUC(%)
RNA-FM	23M	72.47	67.19	79.68
Random	0.2M	69.65	62.14	75.97
Top I (w/o ft)	0.2M	70.62	63.40	**77.16**
CCI (w/o ft)	0.2M	69.10	61.33	75.45
Top I	0.2M	**71.25**	**63.63**	76.97
CCI	0.2M	69.46	62.10	76.04

(d) RBP–RNA Interaction Prediction.

Table 2: Performance comparison of different coreset selection strategies across four structure and interaction prediction tasks for RNA understanding. **Bold** denotes the best results and underline denotes the second-best results.

Methods	Data Size	Bin	SS	Aff	
		ACC(%)	ACC(%)	MAE ↓	RMSE ↓
ESM-C	2.78B	91.63	86.10	1.92	2.44
Random	2M	75.76	67.20	2.39	2.87
Random	0.2M	73.64	66.18	2.51	3.01
Top I	0.2M	77.13	69.34	**2.06**	**2.64**
CCI	0.2M	**79.25**	**71.48**	2.14	2.69

Table 3: Performance of different coresets across three different downstream prediction tasks (Binary Localization Prediction, Secondary Structure Prediction, and PPI Affinity Prediction). **Bold** denotes the best results and underline denotes the second-best results.

over, this observation echoes our results on RNA-FM and suggests that the protein sequence dataset also exhibits a high level of redundancy, further underscoring the potential of data pruning. Although the models pretrained on Top I or CCI coresets still exhibit a performance gap compared to the original ESM-C, this discrepancy can be attributed to the substantial gap in data scale (0.2M vs. 2.78B). Due to current resource constraints, we leave the exploration of more suitable coreset sizes for protein foundation models as future work. Nevertheless, these results still provide encouraging evidence that influence-guided data pruning holds promise across both RNA and protein domains, even under extremely limited data budgets.

6 Conclusion

In this work, we investigate the problem of data pruning for pretraining biological foundation models (BioFMs) at scale, aiming to alleviate the substantial computational demands posed by large-scale pretraining. To this end, we introduce a post-hoc influence-guided data pruning framework that incorporates two complementary selection strategies—Top-k Influence (Top I) and Coverage-Centric Influence (CCI)—to enable scalable and effective coreset construction. Our experiments demonstrate that the proposed framework consistently outperforms random selection across both RNA and protein domains, while in RNA structure prediction tasks, it even achieves performance comparable to the original RNA-FM using less than 1% of the full 23-million-sequence training set. This demonstrates the effectiveness of our data pruning framework. Looking forward, our work offers a promising pathway toward training high-performing BioFMs on compact yet informative subsets, which can facilitate more reproducible, accessible, and sustainable biological AI research.

Acknowledgments

This work was supported by the Major Project of Guangzhou National Laboratory (Grant No. GZNL2024A01003, GZNL2023A02007, GZNL2025C02028 to Jiao Yuan), National Natural Science Foundation of China (Grant No. 32400547 to Jiao Yuan), Pearl River Talent Recruitment Program (2023QN10Y296 to Jiao Yuan), Guangzhou Young Top Talent Program, National Key R&D Program of China (2023YFF1204701 to Jiao Yuan), the Chinese University of Hong Kong (CUHK; award numbers 4937025, 4937026, 5501517 and 5501329 to Yu Li), Shenzhen Medical Research Fund (Grant No. A2503002 to Yu Li), the IdeaBooster Fund (IDBF23ENG05 and IDBF24ENG06 to Yu Li), partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Hong Kong SAR), China

(project no. CUHK 24204023 to Yu Li), a grant from the Innovation and Technology Commission of the Hong Kong SAR, China (project no. GHP/065/21SZ and ITS/247/23FP to Yu Li), and the Research Matching Grant Scheme at CUHK (award numbers 8601603 and 8601663 to Yu Li) from the Research Grants Council, Hong Kong SAR, China.

References

Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; and Winther, O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21): 3387–3395.

Amin, N.; McGrath, A.; and Chen, Y.-P. P. 2019. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence*, 1(5): 246–256.

Brixi, G.; Durrant, M. G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G. A.; King, S. H.; Li, D. B.; Merchant, A. T.; Naghipourfar, M.; Nguyen, E.; Ricci-Tam, C.; Romero, D. W.; Sun, G.; Taghibakshi, A.; Vorontsov, A.; Yang, B.; Deng, M.; Gorton, L.; Nguyen, N.; Wang, N. K.; Adams, E.; Baccus, S. A.; Dillmann, S.; Ermon, S.; Guo, D.; Ilango, R.; Janik, K.; Lu, A. X.; Mehta, R.; Mofrad, M. R.; Ng, M. Y.; Pannu, J.; Re, C.; Schmok, J. C.; St. John, J.; Sullivan, J.; Zhu, K.; Zynda, G.; Balsam, D.; Collison, P.; Costa, A. B.; Hernandez-Boussard, T.; Ho, E.; Liu, M.-Y.; McGrath, T.; Powell, K.; Burke, D. P.; Goodarzi, H.; Hsu, P. D.; and Hie, B. 2025. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*.

Chen, H.; Dong, Y.; Wei, Z.; Huang, Y.; Zhang, Y.; Su, H.; and Zhu, J. 2025. Understanding Pretraining and Fine-tuning from Loss Landscape Perspectives. arXiv:2505.17646.

Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T.; et al. 2022. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*.

Chuai, G.; Ma, H.; Yan, J.; Chen, M.; Hong, N.; Xue, D.; Zhou, C.; Zhu, C.; Chen, K.; Duan, B.; et al. 2018. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome biology*, 19: 1–18.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2020. Selection via Proxy: Efficient Data Selection for Deep Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Danaee, P.; Rouches, M.; Wiley, M.; Deng, D.; Huang, L.; and Hendrix, D. 2018. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic acids research*, 46(11): 5381–5394.

Diddee, H.; and Ippolito, D. 2025. Chasing Random: Investigating the "Gains" Achieved through Instruction Selection Strategies at Scale. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 1943–1957. Mexico City, Mexico: Association for Computational Linguistics.

Duan, H.-C.; Wang, Y.; and Jia, G. 2019. Dynamic and reversible RNA N6-methyladenosine methylation. *Wiley Interdisciplinary Reviews: RNA*, 10(1): e1507.

Guo, Y.; Yu, L.; Wen, Z.; and Li, M. 2008. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9): 3025–3030.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346): 383–393.

Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; et al. 2025. Simulating 500 million years of evolution with a language model. *Science*, eads0018.

He, M.; Yang, S.; Huang, T.; and Zhao, B. 2024. Large-scale Dataset Pruning with Dynamic Uncertainty. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, 7713–7722. IEEE.

Kim, S.; Kim, K.; and Yang, E. 2023. GEX: A flexible method for approximating influence via Geometric Ensemble. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Soenderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. 2019. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6): 520–527.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

Martens, J.; and Grosse, R. B. 2015. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In Bach, F. R.; and Blei, D. M., eds., *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 2408–2417. JMLR.org.

Moal, I. H.; and Fernández-Recio, J. 2012. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20): 2600–2607.

Moser, B. B.; Shanbhag, A. S.; Frolov, S.; Raue, F.; Folz, J.; and Dengel, A. 2025. A Coreset Selection of Core-

set Selection Literature: Introduction and Recent Advances. arXiv:2505.17799.

Pascanu, R.; and Bengio, Y. 2014. Revisiting Natural Gradient for Deep Networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Phillips, J. M. 2017. Coresets and sketches. In *Handbook of discrete and computational geometry*, 1269–1288. Chapman and Hall/CRC.

Pleiss, G.; Zhang, T.; Elenberg, E. R.; and Weinberger, K. Q. 2020. Identifying Mislabeled Data using the Area Under the Margin Ranking. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ren, Y.; Chen, Z.; Qiao, L.; Jing, H.; Cai, Y.; Xu, S.; Ye, P.; Ma, X.; Sun, S.; Yan, H.; Yuan, D.; Ouyang, W.; and Liu, X. 2024. BEACON: Benchmark for Comprehensive RNA Tasks and Language Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Roux, N. L.; Manzagol, P.; and Bengio, Y. 2007. Topmoumoute Online Natural Gradient Algorithm. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 849–856. Curran Associates, Inc.

Sarkisyan, K. S.; Bolotin, D. A.; Meer, M. V.; Usmanova, D. R.; Mishin, A. S.; Sharonov, G. V.; Ivankov, D. N.; Bozhanova, N. G.; Baranov, M. S.; Soylemez, O.; et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603): 397–401.

Schaul, T.; Zhang, S.; and LeCun, Y. 2013. No more pesky learning rates. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, 343–351. JMLR.org.

Shen, T.; Hu, Z.; Sun, S.; Liu, D.; Wong, F.; Wang, J.; Chen, J.; Wang, Y.; Hong, L.; Xiao, J.; et al. 2024. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, 1–12.

Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; and Morcos, A. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; and Wu, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10): 1282–1288.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 9275–9293. Association for Computational Linguistics.

Team, B. A. A.; Chen, X.; Zhang, Y.; Lu, C.; Ma, W.; Guan, J.; Gong, C.; Yang, J.; Zhang, H.; Zhang, K.; et al. 2025. Protenix-Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv*, 2025–01.

Van der Vaart, A. W. 2000. *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, Z.; Wang, Z.; Jiang, J.; Chen, P.; Shi, X.; and Li, Y. 2025. Large language models in bioinformatics: A survey. *arXiv preprint arXiv:2503.04490*.

Xia, X.; Liu, J.; Yu, J.; Shen, X.; Han, B.; and Liu, T. 2023. Moderate Coreset: A Universal Method of Data Selection for Real-world Data-efficient Deep Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xu, M.; Zhang, Z.; Lu, J.; Zhu, Z.; Zhang, Y.; Chang, M.; Liu, R.; and Tang, J. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35: 35156–35173.

Xu, Y.; Zhu, J.; Huang, W.; Xu, K.; Yang, R.; Zhang, Q. C.; and Sun, L. 2023. PrismNet: predicting protein–RNA interaction using in vivo RNA structural information. *Nucleic Acids Research*, 51(W1): W468–W477.

Yang, Y.; Mishra, S.; Chiang, J. N.; and Mirzasoleiman, B. 2024. SmallToLarge (S2L): Scalable Data Selection for Fine-tuning Large Language Models by Summarizing Training Trajectories of Small Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ye, X.; Wu, Y.; Zhang, W.; Jin, C.; and Chen, Y. 2025. Towards Robust Influence Functions with Flat Validation Minima. In *Proceedings of the 42th International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR.

Zhang, J.; Qin, Y.; Pi, R.; Zhang, W.; Pan, R.; and Zhang, T. 2025. TAGCOS: Task-agnostic Gradient Clustered Coreset Selection for Instruction Tuning Data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 4686–4701. Mexico City, Mexico: Association for Computational Linguistics.