On the Importance of Pretraining Data Alignment for Atomic Property Prediction

Anonymous authors
Paper under double-blind review

Abstract

This paper challenges the recent paradigm in atomic property prediction that links progress to growing dataset sizes and computational resources. We show that pretraining on a carefully selected task-aligned dataset can match or even surpass large-scale pretraining, while using only **1/24th of the computational cost**. We introduce the Chemical Similarity Index (CSI), a simple metric for molecular graphs inspired by the Fréchet Inception Distance in computer vision, which quantifies the alignment between upstream pretraining datasets and downstream tasks. By selecting the most aligned dataset with minimal CSI distance, we show that models pretrained on a smaller, focused dataset consistently achieve better performance on downstream tasks than those pretrained on massive, mixed datasets such as JMP. This holds even when the mixed dataset includes the upstream dataset most aligned with the downstream task. Counterintuitively, we also find that indiscriminately adding more data can degrade model performance when the additional data is poorly aligned with the target task. Our findings highlight that **quality often outperforms quantity** in pretraining for atomic property prediction.

1 Introduction

Machine learning is transforming molecular modeling, driving advancements in accurate predictions and simulations of molecular behavior (Chanussot et al., 2021; Tran et al., 2023; Liao et al., 2023). These breakthroughs directly impact the acceleration of progress in crucial fields such as drug discovery (Huang et al., 2021) and global climate change mitigation (Sriram et al., 2024). The improvements in this field have been primarily attributed to innovations in model architectures (Liao et al., 2023; Gasteiger et al., 2021; Passaro & Zitnick, 2023) and the growing availability of large-scale molecular datasets. In recent years, the sizes of molecular datasets have increased dramatically - from tens of thousands of examples (Christensen & Von Lilienfeld, 2020; Chmiela et al., 2023; Wu et al., 2018) to hundreds of millions (Chanussot et al., 2021; Tran et al., 2023). This rapid growth in scale has also caused a surge in the computational resources required for pretraining, increasing from a few days on a single GPU to over a thousand GPU-days (Shoghi et al., 2023; Liao et al., 2023). This trend begs the question:

💡 ***Is scaling data and resources the only path forward in atomic property prediction, or can intelligent data selection achieve similar performance more efficiently?***

While data selection strategies for pretraining have been explored in fields like natural language processing (Penedo et al., 2024) and computer vision (Hammoud et al., 2024; Li et al., 2023), this area remains largely underexplored in atomic property prediction, where unique challenges arise. In this paper, we challenge the prevailing assumption that "bigger is better" by exploring whether a smaller, strategically selected dataset can lead to comparable or even superior performance while substantially reducing computational demands. We introduce a pretraining paradigm that shifts the focus from data and compute scaling to selecting the most relevant upstream dataset for improved downstream performance.

Through a simple baseline, our experiments reveal two key insights: **(1) Competitive Performance Can Be Achieved with 24× Fewer Resources:** Selecting upstream datasets based on their alignment with

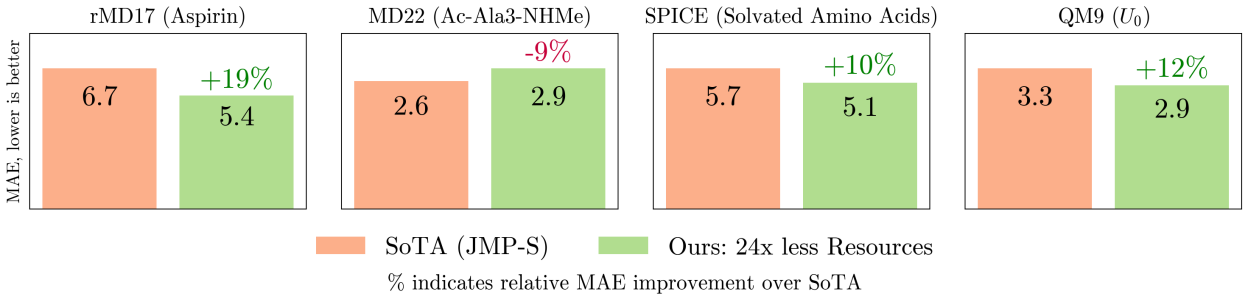

Figure 1: **Pretraining on a High-Quality, Task-Aligned Dataset.** Pretraining on a carefully selected high-quality dataset achieves comparable or superior mean absolute error (MAE) across tasks while reducing computational cost by a factor of 24 compared to JMP-S, which is pretrained on all upstream datasets. Lower MAE indicates better performance.

the downstream task achieves performance on par with or exceeding that of large-scale pretrained models like JMP (Shoghi et al., 2023) while utilizing only **1/24th** of the computational resources, as shown in Figure 1. **(2) Quality Outperforms Quantity:** Expanding the pretraining dataset by incorporating additional data from less aligned sources can negatively impact downstream performance rather than enhance it.

To explore the potential of dataset selection for pretraining in atomic property prediction, we introduce the Chemical Similarity Index (CSI), a simple metric inspired by the Fréchet Inception Distance (FID) from computer vision. CSI measures the alignment between an upstream dataset and a downstream task, enabling the selection of chemically relevant pretraining data. By focusing on these highly relevant datasets, we significantly reduce computational costs while maintaining competitive performance and, in many cases, achieving improvements. While large-scale datasets like OC20 (Chanussot et al., 2021; Tran et al., 2023) and mixed datasets like JMP (Shoghi et al., 2023) are popular choices for pretraining in molecular domains (Kolluru et al., 2022; Shoghi et al., 2023), our findings challenge their universal utility. Surprisingly, pretraining on a single, carefully selected dataset guided by CSI often outperforms models trained on mixtures, even when those include the most relevant dataset.

The contributions of this paper are threefold: (1) We introduce a novel framework for computationally efficient pretraining of molecular machine learning models, demonstrating that strategic data selection can match or outperform models trained on much larger datasets. (2) We propose the Chemical Similarity Index (CSI), a metric for assessing the alignment between upstream and downstream molecular datasets, enabling effective dataset selection. (3) We provide an extensive empirical evaluation demonstrating the effectiveness of our approach, offering a practical and efficient alternative to the current trend of ever-increasing data and computational costs in molecular machine learning.

2 Related Work

Pretraining for Atomic Property Prediction. Inspired by the success of pretraining in computer vision and natural language processing, pretraining for atomic property prediction has gained significant attention in recent years. Most approaches in molecular machine learning focus on self-supervised learning (Liu et al., 2021; Jiao et al., 2023; Chen et al., 2021; Kolluru et al., 2022; Zhou et al., 2022a; Ji et al., 2024), as generating labels for molecular datasets is computationally expensive. In contrast, fewer studies explore the effectiveness of supervised transfer learning (Smith et al., 2019; 2018; Kolluru et al., 2022). However, in both self-supervised and supervised settings, the focus has primarily been on improving feature representation, often overlooking the impact of pretraining dataset relevance on downstream performance. Recently, Shoghi et al. (2023) introduced Joint Multi-domain Pre-training (JMP), enabling pretraining on various upstream sources concurrently. While effective, JMP pretraining requires enormous computational resources to reproduce and does not reveal how each upstream source impacts downstream performance. Our work addresses this gap by systematically studying the relationship between upstream pretraining datasets and downstream

performance, enabling researchers to develop effective pretraining models even with limited computational resources.

Computational Budgeting. Recent research highlights the importance of studying model performance under computationally budgeted setups. In continual learning (CL), works by Prabhu et al. (2023) and Ghunaim et al. (2023) show that simple baselines often outperform state-of-the-art methods in compute-constrained settings. TiC-CLIP (Garg et al., 2024) further demonstrates efficient rehearsal-based training for time-continuous data. For Vision Transformers, Pan et al. (2022) propose a framework to dynamically control model complexity during training, achieving competitive performance under varying budgets. Li et al. (2019) formalize budgeted training, showing that budget-aware learning rate schedules, such as linear decay, are critical for robust performance across tasks like image classification and object detection. In multi-domain learning, Berriel et al. (2019) introduce Budget-Aware Adapters, which reduce computational complexity while maintaining accuracy by selecting relevant feature channels. These findings across domains emphasize the critical need for more efficient approaches that can achieve competitive performance while minimizing computational costs.

Data Selection. Efficient training through data selection has been explored via two primary approaches: subset selection and dataset distillation. Subset selection aims to identify a representative subset of the training data that matches or even outperforms training on the full dataset. Several methods have been proposed for vision and NLP tasks (Attendu & Corbeil, 2023; Killamsetty et al., 2021a;b; Kaushal et al., 2019; Bairi et al., 2015; Lapedriza et al., 2013). Dataset distillation, introduced by Wang et al. (2018), focuses on generating a smaller, synthetic subset of the dataset that preserves performance while reducing training time and storage requirements. Subsequent work has explored techniques such as meta-learning (Zhou et al., 2022b; Nguyen et al., 2021a;b), gradient matching (Zhao et al., 2021), and distribution matching (Zhao & Bilen, 2023). While most research in distillation has focused on vision tasks, a few studies have extended it to graph data (Jin et al., 2022b; Liu et al., 2022; Jin et al., 2022a), though primarily targeting knowledge and social graphs rather than molecular graphs.

Two recent vision studies are particularly relevant to our work. First, Hammoud et al. (2024) shows that increasing pretraining data diversity enhances performance only when distribution shifts between upstream and downstream tasks are minimized. Second, Li et al. (2023) introduces a method to dynamically leverage the open web, reducing the distribution gap between upstream and downstream tasks through targeted representation learning. Findings from other domains suggest that aligning upstream datasets may be crucial for effective pretraining.

Comparison to Our Work. To the best of our knowledge, no prior work has specifically explored upstream dataset selection for molecular graphs, which present unique challenges due to their structural and chemical complexity. In this work, we take the first step in addressing this gap by focusing on aligning upstream and downstream distributions at the dataset level rather than subselecting at a sample-wise level or creating a synthetic distilled version of the dataset.

3 Formulation and Setup

In this section, we present our problem setup, notion of a computational budget, and the formulation of dataset similarity. We then detail how we adapt the Fréchet Inception Distance (FID) to the molecular domain, yielding the *Chemical Similarity Index (CSI)*. Our setup is illustrated in Figure 2. Throughout this work, we use the term 'molecular' broadly to encompass both molecular and materials domains, as well as their respective datasets.

3.1 Formal Setting

Upstream and Downstream Datasets. Let $\{\mathcal{D}_u^{(1)}, \mathcal{D}_u^{(2)}, \ldots, \mathcal{D}_u^{(K)}\}$ denote a collection of K *upstream* (pretraining) datasets, each containing molecular structures paired with relevant atomic properties (e.g., energies and forces). In the typical paradigm, upstream datasets are typically aggregated into a single

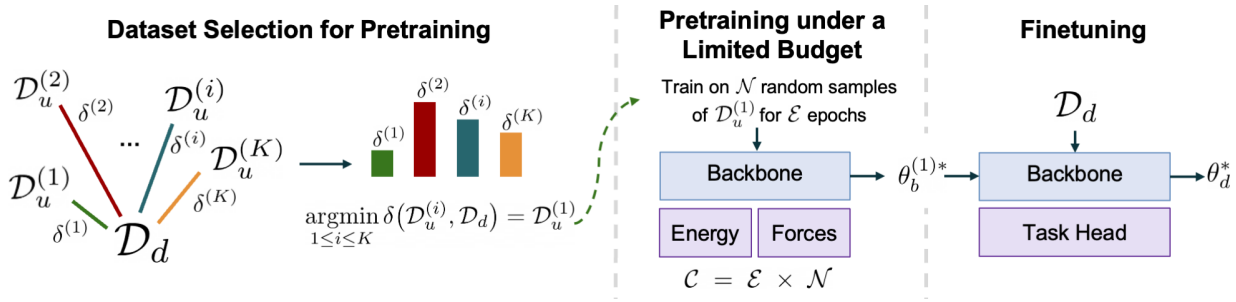

Figure 2: **Pipeline Overview**. Our paradigm for pretraining and finetuning consists of two new components: (1) *Dataset Selection Stage*, where a distance metric δ is employed to identify the dataset that is most similar to our downstream task dataset $\mathcal{D}_d$, in this case $\mathcal{D}_u^{(1)}$. This selected dataset is then used for pretraining the model. (2) *Limited Budget Pretraining*, where we impose a training budget by subsampling $\mathcal{N}$ random samples from $\mathcal{D}_u^{(1)}$ and training the model for $\mathcal{E}$ epochs. This results in a computational budget of $\mathcal{C} = \mathcal{E} \times \mathcal{N}$. The pretrained backbone $\theta_b^{(1)*}$ is subsequently finetuned on the downstream task dataset $\mathcal{D}_d$ to obtain the final model parameters θ_d^*.

pretraining set:

$$\mathcal{D}_u = \bigcup_{i=1}^{K} \mathcal{D}_u^{(i)}. \tag{1}$$

We further define $\mathcal{D}_d$ as the *downstream* dataset, which focuses on a specific prediction task (e.g., predicting an atomic property).

Multi-task Pretraining. We consider a neural network $\Phi(\cdot;\theta)$, where θ encompasses the shared backbone parameters θ_b and task-specific head parameters θ_e (for energy prediction) and θ_f (for force prediction). During pretraining, the network is trained to simultaneously predict energies and forces. Formally, the multi-task pretraining objective over an upstream dataset $\mathcal{D}_u^{(i)}$ is given by:

$$\theta^{(i)*} \;=\; \arg\min_{\theta} \; \mathcal{L}_{\text{pretrain}}(\theta; \mathcal{D}_u^{(i)}), \tag{2}$$

where $\theta = \{\theta_b, \theta_e, \theta_f\}$ and

$$\mathcal{L}_{\text{pretrain}}(\theta; \mathcal{D}_u^{(i)}) \;=\; \alpha\, \mathcal{L}_{\text{energy}}(\theta_b, \theta_e; \mathcal{D}_u^{(i)}) \;+\; \beta\, \mathcal{L}_{\text{forces}}(\theta_b, \theta_f; \mathcal{D}_u^{(i)}). \tag{3}$$

We compute $\mathcal{L}_{\text{energy}}$ using the Mean Absolute Error (MAE) and $\mathcal{L}_{\text{forces}}$ using the mean per-atom Euclidean (L2) distance. Coefficients α and β weight the importance of energy and force tasks, respectively. Following the JMP paper (Shoghi et al., 2023), we set $\beta > \alpha$ to prioritize accurate force predictions in atomistic modeling. Pretraining can be performed on either the joint upstream dataset $\mathcal{D}_u$, similar to JMP (Shoghi et al., 2023), or on an individual upstream dataset $\mathcal{D}_u^{(i)}$, as in our selective setting.

Fine-Tuning. After the multi-task pretraining phase, the task-specific heads θ_e and θ_f are discarded, and a new task-specific head θ_h is attached to the pretrained backbone θ_b. The downstream objective then becomes:

$$\theta_d^* \;=\; \arg\min_{\theta_b, \theta_h} \; \mathcal{L}_{\text{finetune}}(\theta_b, \theta_h; \theta_b^{(i)*}, \mathcal{D}_d), \tag{4}$$

where $\theta_b^{(i)*}$ denotes the pretrained backbone parameters from Eq. (2). Intuitively, the downstream training refines the shared backbone parameters θ_b and learns the task-specific head θ_h to capture the target property in $\mathcal{D}_d$.

In this paper, we consider two additional needed definitions for this setting: (1) computational budget and (2) dataset similarity.

Computational Budget. Following Hammoud et al. (2024), we define the *computational budget* $\mathcal{C}$ to be the product of the number of epochs $\mathcal{E}$ and the number of unique samples $\mathcal{N}$ in the pretraining dataset:

$$\mathcal{C} \;=\; \mathcal{E} \;\times\; \mathcal{N}. \tag{5}$$

Hence, the computational budget $\mathcal{C}$ represents the total number of samples processed over training. It naturally splits into two factors: the dataset size ($\mathcal{N}$) and the number of passes through it ($\mathcal{E}$). The choice of $\mathcal{C}$ depends on the available computing resources. In our main experiments, we fix $\mathcal{C}$, $\mathcal{N}$, and $\mathcal{E}$ to ensure a fair comparison across different upstream datasets. We also include experiments in which $\mathcal{N}$ (and thus $\mathcal{C}$) varies, in order to analyze the impact of dataset size and total compute on downstream performance.

Dataset Similarity. A key objective of this work is to estimate how well an upstream dataset $\mathcal{D}_u$ aligns with a downstream dataset $\mathcal{D}_d$. We therefore seek a distance metric

$$\delta(\mathcal{D}_u, \mathcal{D}_d)$$

that quantifies their alignment or "similarity." In principle, a lower value of $\delta(\mathcal{D}_u, \mathcal{D}_d)$ reflects a higher degree of alignment between the upstream and downstream distributions. Thus, among multiple candidate upstream datasets $\{\mathcal{D}_u^{(1)}, \ldots, \mathcal{D}_u^{(K)}\}$, the one that minimizes

$$\operatorname*{argmin}_{1 \le i \le K} \delta\big(\mathcal{D}_u^{(i)}, \mathcal{D}_d\big)$$

should provide the most effective pretraining for $\mathcal{D}_d$. In this paper, we empirically test this assumption, examining whether lower δ-values indeed correlate with improved downstream performance. Motivated by this, we use δ as a principled metric to guide dataset selection for Eq. (1). Instead of aggregating all upstream datasets, we modify the pretraining setup to use only the single dataset $\mathcal{D}_u^{(i)}$ that best aligns with the downstream task under a fixed computational budget.

3.2 The Chemical Similarity Index (CSI)

Recap of FID. Our proposed *Chemical Similarity Index (CSI)* draws its inspiration from the well-known Fréchet Inception Distance (FID) (Heusel et al., 2017). Recall that FID is commonly used in computer vision to compare two sets of images via their feature distributions. Specifically, if one extracts features (e.g., from an Inception network) for datasets X and Y and denotes their empirical means and covariances by μ_X, Σ_X and μ_Y, Σ_Y, then

$$\mathrm{FID}(X, Y) \;=\; \|\mu_X - \mu_Y\|^2 \;+\; \mathrm{Tr}\Big(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{1/2}\Big). \tag{6}$$

The central idea is to represent each sample in a feature space where distances encode semantic similarity and then compare the distributions of these representations for the two datasets.

To adapt FID for graph-structured molecular data, we compute the CSI metric using node embeddings as features and apply class-balanced sampling to ensure representative coverage of molecular types in each upstream dataset. For computational feasibility, we subsample 10k instances from both the upstream and downstream datasets. To keep the metric independent of the baselines evaluated in this study, we extract features using EquiformerV2 (Liao et al., 2023) pretrained on OC20 (Chanussot et al., 2021).

CSI Between Upstream and Downstream Results. In Figure 3, we present the CSI values for pairs of upstream and downstream tasks related to energy and force predictions, with additional details about the datasets and targets provided in Section 4. ANI-1x (Smith et al., 2020) consistently achieves the closest alignment across all downstream tasks, reflecting its design goal of maximizing chemical diversity. Transition-1x (Schreiner et al., 2022), which focuses on transition states, shows as the second most aligned dataset, suggesting that its emphasis on high-energy transition states leads to partial overlap with downstream distributions. In contrast, the catalysis datasets, OC20 (Chanussot et al., 2021) and OC22 (Tran et al., 2023), exhibit the weakest alignment. While OC20 and OC22 are often favored for pretraining (Shoghi et al., 2023; Kolluru et al., 2022) due to their scale and chemical diversity, our metric suggests they may not align well with the considered downstream tasks. Next, we examine whether these alignment values correlate with downstream performance.

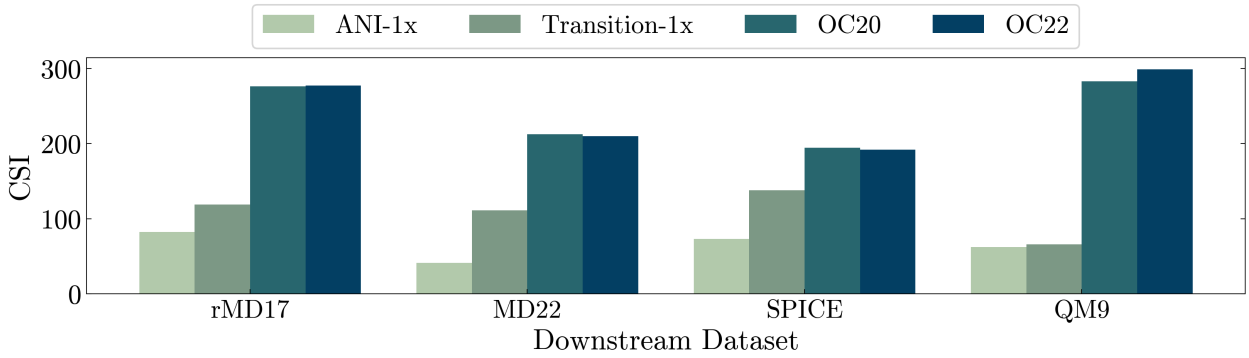

Figure 3: **Alignment Between Upstream and Downstream Using CSI.** We assess how well the extracted representations from each upstream dataset align with downstream tasks using our CSI metric, where lower values indicate stronger alignment. ANI-1x demonstrates the closest feature alignment with downstream tasks, whereas OC20 and OC22 show the weakest alignment.

4 Experiments

We evaluate the impact of pretraining on different upstream datasets for downstream performance and investigate how well the CSI values in Figure 3 reflect the relevance of these datasets. We begin by defining the datasets, baselines, and evaluation setup.

Upstream Datasets: Following JMP (Shoghi et al., 2023), we perform pretraining on upstream datasets of small molecules, including ANI-1x (Smith et al., 2020) and Transition-1x (Schreiner et al., 2022), as well as large-scale catalysis datasets, OC20 (Chanussot et al., 2021) and OC22 (Tran et al., 2023). These datasets vary in domain focus and graph size, enabling us to examine how these factors impact the generalization of pretraining across downstream tasks. The ground-truth labels, energy and forces, are computed using Density Functional Theory (DFT).

Downstream Datasets: For downstream evaluation, we focus on in-distribution (ID) tasks involving energy or force prediction, following the definition in JMP (Shoghi et al., 2023). We discuss out-of-distribution (OOD) tasks in Section 5. Given the large number of pairs for ID evaluation, we focus on the first molecule for force tasks and the energy target for the multi-property dataset QM9 (Wu et al., 2018). The selected targets and their corresponding datasets are: Aspirin in rMD17 (Christensen & Von Lilienfeld, 2020), Ac-Ala3-NHMe in MD22 (Chmiela et al., 2023), solvated amino acids in SPICE (Eastman et al., 2023), and U_0 in QM9 (Wu et al., 2018).

Baselines: We report the original performance of JMP, where "JMP-S" and "JMP-L" correspond to the small and large backbones, respectively. Additionally, we present our reproduced fine-tuning results using the official JMP checkpoints, denoted as "JMP-S*" and "JMP-L*".

For our budgeted evaluation, we present results in two categories: pretraining on a single upstream dataset and pretraining on a joint combination of all upstream datasets. For single-dataset experiments, we randomly sample $\mathcal{N}$ instances from the original upstream data. For joint pretraining, we construct the training set using two different strategies. (1) Balanced Sampling, where an equal number of samples is drawn from each of the four upstream datasets, totaling $\mathcal{N}$ samples; and (2) Temperature-Based Sampling, which preserves the dataset proportions used in the full 120M sample set of JMP (Shoghi et al., 2023).

Evaluation Setup: We pretrain the GemNet-OC-S model (Gasteiger et al., 2022) on each individual upstream dataset, as well as on joint configurations that combine all upstream datasets, following the baseline setups. For our main experiments, we set a fixed computational budget of $\mathcal{C} = 10$M, achieved by training on $\mathcal{N} = 2$M samples for $\mathcal{E} = 5$ epochs. This budget ensures accessibility and reproducibility, with each pretraining run completing within 1 to 2 days on an A100 GPU. This represents a 24× reduction in computational cost compared to the pretraining budget used in JMP (Shoghi et al., 2023). Additional budget

Table 1: **In-Distribution Evaluation for energy and force targets.** We report test MAE when fine-tuning on downstream targets, as detailed in Downstream Datasets (Section 4). The top section represents models pretrained with the large-scale JMP budget, while the lower two sections show results under a limited budget. JMP-S* denotes reproduced results.

$\mathcal{C}$	**Upstream Data**	**Backbone**	**rMD17** (meV/Å)	**MD22** (meV/Å)	**SPICE** (meV/Å)	**QM9** (meV)
240M	Joint (Temperature)	JMP-L (GemNet-OC-L)	5.1	1.92	4.75	2.9
		JMP-S (GemNet-OC-S)	6.7	2.64	5.71	3.3
		JMP-S*(GemNet-OC-S)	6.8	3.21	5.60	3.4
10M	ANI-1x	GemNet-OC-S	**5.4**	**2.90**	**5.13**	**2.9**
	Transition-1x		10.1	3.73	7.55	3.2
	OC20		14.6	4.53	8.74	4.8
	OC22		16.0	5.20	10.73	5.7
10M	Joint (Balanced)	GemNet-OC-S	9.4	3.62	7.02	3.2
	Joint (Temperature)		11.0	4.03	7.98	3.6

configurations are explored in later sections and the appendix. Each pretrained model is then fine-tuned separately on each downstream task.

4.1 Does CSI Correlate with Better Performance?

In Figure 3, we presented CSI values quantifying the alignment between each upstream and downstream dataset. The results indicate that ANI-1x exhibits the highest alignment to all downstream datasets. This finding raises a critical question:

Can CSI reliably guide the selection of pretraining datasets to achieve optimal performance on specific downstream tasks?

Table 1 summarizes the downstream performance of models pretrained on different datasets in the in-distribution setting. Surprisingly, the model pretrained on ANI-1x, despite being the smallest dataset, consistently outperforms all other individual datasets as well as the joint variants. For instance, on the rMD17, SPICE, and QM9 datasets, the model pretrained on ANI-1x achieves MAEs of 5.4, 5.13, and 2.9, compared to 6.7, 5.71, and 3.3 for JMP-S. This strong performance is achieved with less than 5% of the pretraining computational budget used by JMP-S. In contrast, and consistent with our CSI scores, pretraining on the least aligned upstream datasets, such as OC20 and OC22, leads to the worst downstream performance compared to ANI-1x.

Furthermore, temperature-based joint training, which follows the JMP formulation and emphasizes high-CSI datasets such as OC20 and OC22, performs worse overall. Balanced joint training offers a slight improvement by including a higher proportion of samples from the more aligned datasets ANI-1x and Transition-1x, yet remains inferior to individual pretraining on ANI-1x. These results suggest that, under a limited budget, mixing upstream datasets with varying CSI values is suboptimal and requires significantly more computational resources to achieve competitive performance.

Takeaway. Our experiments reveal three key insights for in-distribution downstream tasks: (1) Task-aligned upstream datasets such as ANI-1x outperform larger joint datasets, even when ANI-1x is part of the joint composition. (2) Joint pretraining can match the benefits of highly relevant pretraining, but it requires significantly more compute and training time. (3) CSI effectively predicts downstream performance, as lower CSI values (e.g., ANI-1x) consistently correlate with better results.

4.2 What is the Effect of Computational Budget?

Building on our earlier findings, we now investigate how varying the computational budget impacts downstream performance. Specifically, we ask:

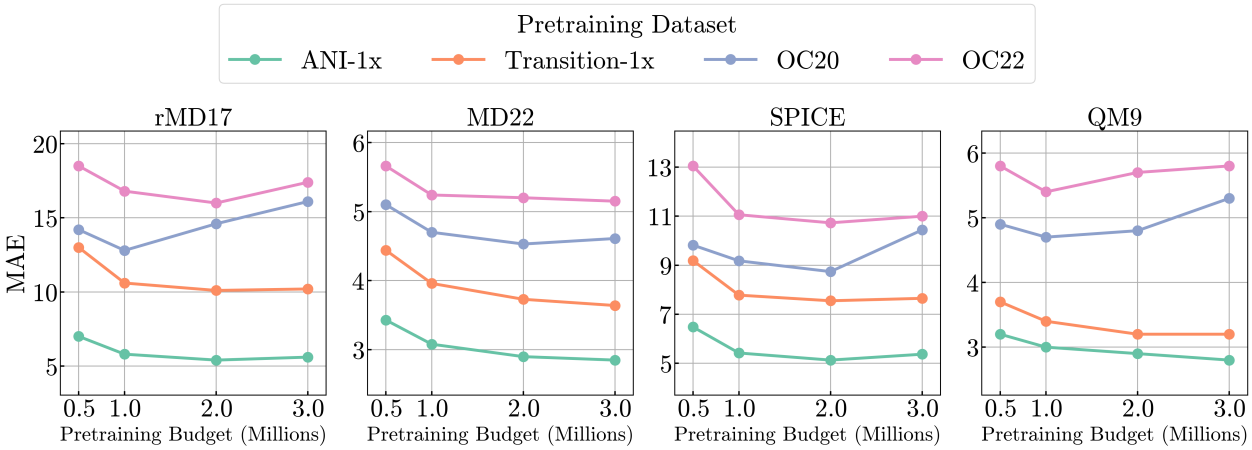

Figure 4: **Effect of Computational Budget on Performance.** While fixing the number of epochs ($\mathcal{E}$) to 5, we vary the number of training samples across $\mathcal{N}$ = 0.5M, 1M, 2M, and 3M. Our findings are consistent across budget levels where the upstream dataset with the lowest CSI yields the best downstream performance.

Do our findings about dataset alignment in terms of CSI hold across different budget levels?

Figure 4 shows the downstream MAE across pretraining budgets of 0.5M, 1M, 2M, and 3M samples (each trained for 5 epochs). Pretraining on ANI-1x consistently yields the best downstream performance across all budget levels, reinforcing the importance of task-aligned, low-CSI datasets. We observe that increasing the budget for ANI-1x beyond 2M tends to offer diminishing returns on rMD17 and SPICE, suggesting slight overfitting. In contrast, increasing the pretraining budget for high-CSI datasets (i.e., OC20 and OC22) often degrades downstream performance more drastically, particularly on rMD17, SPICE, and QM9. These results highlight that allocating more compute to misaligned upstream tasks can reduce the model's ability to generalize to downstream tasks.

Takeaway. Our findings are consistent across budget levels: the upstream dataset with the lowest CSI yields the best downstream performance.

4.3 What is the Effect of Changing the Backbone Size?

In the previous sections, we used the small variant, GemNet-OC-S, as our backbone. Here, we address the question:

Does the correlation between CSI and downstream performance hold across different backbone sizes?

Table 2 reports the downstream performance using the large variant, GemNet-OC-L, as the backbone. We also include our best attempt at reproducing the baseline results using JMP-L pretraining (denoted as "JMP-L*").

Consistent with the results on the small backbone, models pretrained on ANI-1x achieve the best performance across all downstream tasks, aligning with its low CSI values. Notably, using a small computation budget of $\mathcal{C}$ = 10M (i.e., 2M samples over 5 epochs), ANI-1x outperforms JMP-L, which was pretrained with $\mathcal{C}$ = 240M on a joint upstream dataset. We obtain state-of-the-art results with an MAE of 4.8 on Aspirin (rMD17) and 2.6 on U_0 (QM9), demonstrating that strong dataset alignment can outweigh large-scale pretraining even with increased model capacity. While larger backbones improve overall performance, the gap between aligned and misaligned upstream datasets persists. High-CSI datasets like OC20 and OC22 still underperform, reaffirming the importance of dataset alignment.

Table 2: **Effect of Changing the Backbone Size.** We analyze the impact of using a larger variant of GemNet-OC and find that, irrespective of backbone size, relevance-based upstream dataset selection consistently outperforms costly large-scale joint pretraining.

$\mathcal{C}$	**Upstream Data**	**Backbone**	**rMD17** (meV/Å)	**MD22** (meV/Å)	**SPICE** (meV/Å)	**QM9** (meV)
240M	Joint (Temperature)	JMP-L (GemNet-OC-L)	5.1	1.92	4.75	2.9
		JMP-L* (GemNet-OC-L)	5.3	2.59	4.91	3.0
10M	ANI-1x	GemNet-OC-L	**4.8**	**2.54**	**5.24**	**2.6**
	Transition-1x		9.7	3.56	7.42	3.0
	OC20		13.8	3.90	9.24	4.6
	OC22		12.0	4.14	10.43	4.0

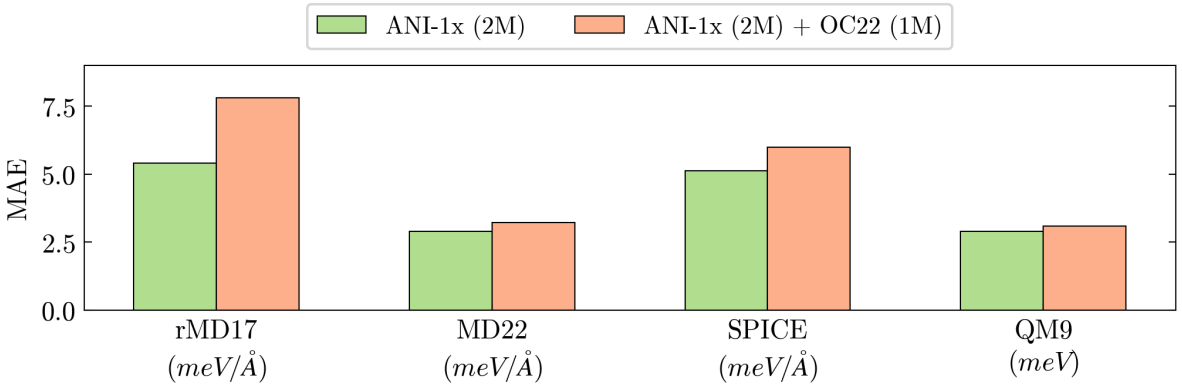

Figure 5: **Impact of Adding Less Aligned Pretraining Data.** Adding $1M$ OC22 samples to a $2M$-sample ANI-1x baseline worsens downstream performance despite a larger pretraining budget. This highlights the importance of dataset alignment and the value of the CSI metric for effective pretraining.

Takeaway. Our findings hold across backbone sizes: scaling up the model does not change the relative utility of upstream datasets. Alignment-based upstream dataset selection outperforms large-scale dataset mixing, even under high-capacity settings and at significantly lower computational budgets.

4.4 Is More Diverse Data Always Better?

A common assumption in pretraining is that larger and more diverse datasets lead to better generalization. This intuition motivates the JMP framework, where a large-scale pretraining budget of $\mathcal{C} = 240$M led to strong downstream results. However, it remains unclear whether this benefit comes from the size, the diversity, or the alignment of the data with the downstream task. Here, we revisit this assumption through a targeted experiment:

Does increasing data diversity by adding less aligned sources improve or harm downstream performance?

To test this, we compare two settings: (1) pretraining on $\mathcal{N} = 2$M unique samples from ANI-1x, the most CSI-aligned dataset, and (2) pretraining on a mixture of 2M ANI-1x samples and 1M OC22 samples (i.e., $\mathcal{N} = 3$M), both trained for 5 epochs. As shown in Figure 5, simply adding OC22 results in worse downstream performance across all four tasks, despite the increase in data volume. This indicates that adding less aligned data may interfere with the knowledge transfer gained from aligned pretraining sources.

Takeaway. Our results challenge the intuitive strategy of adding diversity to pretraining datasets without considering alignment. CSI provides a practical signal for curating upstream data that supports better generalization, especially under constrained budgets.

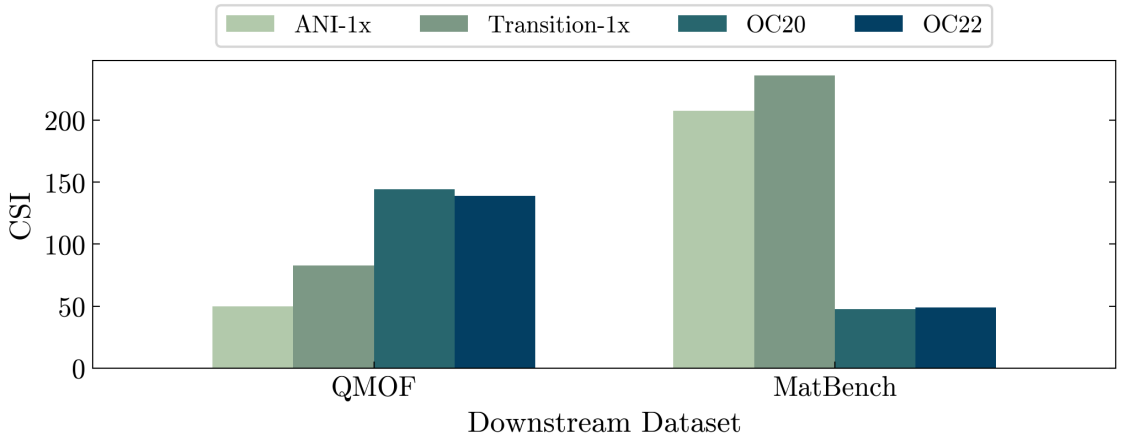

Figure 6: **CSI Between Upstream and OOD Downstream Tasks.** CSI values predict that ANI-1x is the best pretraining choice for QMOF, while OC20 and OC22 are best for MatBench.

Table 3: **OOD Task Performance Across Upstream Sources.** We compare the CSI-predicted best upstream sources with actual downstream performance on OOD tasks (QMOF, MatBench, and QM9's Δ_ϵ). While CSI aligns well with QM9's OOD label, it mispredicts the best source for MatBench. Joint pretraining generally improves performance, highlighting the benefits of diverse upstream sources for OOD generalization.

$\mathcal{C}$	**Upstream Data**	**Backbone**	**QM9** [Δ_ϵ] (meV)	**QMOF** (eV)	**MatBench** [fold0 / mean] (cm^{-1})
240M	Joint (Temperature)	JMP-S (GemNet-OC-S)	23.1	0.18	26.60 / 22.77
		JMP-S* (GemNet-OC-S)	24.0	0.19	24.77 / 21.48
10M	ANI-1x	GemNet-OC-S	**24.5**	0.22	30.09 / 29.60
	Transition-1x		25.3	0.22	52.22 / 38.56
	OC20		30.8	0.22	37.52 / 30.88
	OC22		35.6	0.22	32.78 / 27.55
10M	Joint (Balanced)	GemNet-OC-S	27.3	**0.21**	**26.11 / 24.87**
	Joint (Temperature)		27.9	**0.21**	26.63 / 25.61

5 Beyond In-Distribution

Recall that our pretraining process is conducted on upstream tasks involving molecules and catalysts, with energy and force as targets. For downstream tasks with different labels (e.g., band gap in QMOF) or from distinct chemical domains such as materials (e.g., MatBench and QMOF), we classify these as out-of-distribution (OOD). While our main results focused on ID evaluation, here we explore our metric's applicability to OOD tasks. Specifically, we examine three cases: the Band Gap property from QMOF (Rosen et al., 2021), Phonons (the first non-energy target in JMP tables) from MatBench (Dunn et al., 2020), and Δ_ϵ from QM9, explicitly categorized as OOD in the JMP paper.

In Figure 6, we present the CSI values for OOD domains, where the OOD label (Δ_ϵ) for QM9 follows the same values as in Figure 3. We observe that QMOF exhibits a pattern similar to other ID domains shown in Figure 3. However, MatBench displays a distinct pattern, showing strong correlation with OC20 and OC22, followed by ANI-1x and Transition-1x. Next, we analyze the correlation between CSI and downstream performance under OOD evaluation.

Table 3 shows that Δ_ϵ in QM9 aligns with the CSI pattern, similar to ID evaluation, suggesting that CSI is effective for OOD in the label space. In QMOF, the different upstream sources achieve similar performance which lags behind the full pretraining by JMP. For MatBench (evaluated over 5 folds), OC22 achieves the

best mean performance while OC20 lags behind, despite our metric predicting both to be equally suitable. Additionally, for both QMOF and MatBench, joint pretraining variants generalize better than individual sources. This suggests that when the downstream domain differs from all upstream sources, mixing diverse upstream domains provides the best performance.

While CSI reliably guides dataset selection for in-distribution tasks, its effectiveness in OOD scenarios is less consistent. This may stem from the limited diversity of the backbone used for feature extraction, which was pretrained only on energy and force targets. Future work could explore using backbones pretrained on broader sets of chemical properties or incorporating more diverse upstream domains to better capture variation across OOD tasks. Another promising direction is to leverage foundation models trained on multi-modal or multi-objective tasks, which may offer more transferable representations for similarity assessment across varied downstream domains.

6 Conclusion

This paper challenges the prevailing trend of scaling data and computational resources in atomic property prediction by demonstrating that strategic data selection based on dataset alignment can achieve comparable or superior performance with significantly fewer resources. We introduce the Chemical Similarity Index (CSI), a simple metric that quantifies the alignment between upstream pretraining datasets and downstream tasks, enabling the selection of high-quality, task-aligned pretraining data. Our experiments reveal that smaller, focused datasets often outperform larger, mixed ones, and that indiscriminately adding data can degrade performance when relevance is low. These findings highlight that alignment, rather than scale alone, is the key to effective pretraining, and they point toward a more principled, efficient, and sustainable direction for future research in atomic property prediction.

References

Jean-Michel Attendu and Jean-Philippe Corbeil. Nlu on data diets: Dynamic data subset selection for nlp classification tasks. 2023.

Ramakrishna Bairi, Rishabh Iyer, Ganesh Ramakrishnan, and Jeff Bilmes. Summarization of multi-document topic hierarchies using submodular mixtures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 553–563, 2015.

Rodrigo Berriel, Stephane Lathuillere, Moin Nabi, Tassilo Klein, Thiago Oliveira-Santos, Nicu Sebe, and Elisa Ricci. Budget-aware adapters for multi-domain learning. In *ICCV*, pp. 382–391, 2019.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Dong Chen, Kaifu Gao, Duc Duy Nguyen, Xin Chen, Yi Jiang, Guo-Wei Wei, and Feng Pan. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature communications*, 12 (1):3521, 2021.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023.

Anders S Christensen and O Anatole Von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4):045018, 2020.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Saurabh Garg, Hadi Pour Ansari, Mehrdad Farajtabar, Sachin Mehta, Raviteja Vemulapalli, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models. In *ICLR*, 2024. URL `https://arxiv.org/abs/2310.16226`.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *NeurIPS*, 34:6790–6802, 2021.

Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. *arXiv preprint arXiv:2204.02782*, 2022.

Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, and Bernard Ghanem. Real-time evaluation in online continual learning: A new hope. In *CVPR*, pp. 11888–11897, 2023.

Hasan Abed Al Kader Hammoud, Tuhin Das, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. On pretraining data diversity for self-supervised learning. 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL `https://openreview.net/forum?id=8nvgnORnoWr`.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In *NeurIPS*, 2024. URL `https://openreview.net/forum?id=64V40K2fDv`.

Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, and Yang Liu. Energy-motivated equivariant pretraining for 3d molecular graphs. In *AAAI*, volume 37, pp. 8096–8104, 2023.

Wei Jin, Xianfeng Tang, Haoming Jiang, Zheng Li, Danqing Zhang, Jiliang Tang, and Bing Yin. Condensing graphs via one-step gradient matching. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 720–730, 2022a.

Wei Jin, Lingxiao Zhao, Shichang Zhang, Yozen Liu, Jiliang Tang, and Neil Shah. Graph condensation for graph neural networks. In *ICLR*, 2022b.

Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *WACV*, pp. 1289–1299. IEEE, 2019.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *ICML*, pp. 5464–5474. PMLR, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI*, volume 35, pp. 8110–8118, 2021b.

Adeesh Kolluru, Nima Shoghi, Muhammed Shuaibi, Siddharth Goyal, Abhishek Das, C Lawrence Zitnick, and Zachary Ulissi. Transfer learning using attentions across atomic systems with graph neural networks (taag). *The Journal of Chemical Physics*, 156(18), 2022.

Agata Lapedriza, Hamed Pirsiavash, Zoya Bylinskii, and Antonio Torralba. Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*, 2013.

Alexander C Li, Ellis Brown, Alexei A Efros, and Deepak Pathak. Internet explorer: Targeted representation learning on the open web. In *ICML*. PMLR, 2023.

Mengtian Li, Ersin Yumer, and Deva Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints. *arXiv preprint arXiv:1905.04753*, 2019.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *ICLR*, 2023.

Mengyang Liu, Shanchuan Li, Xinshi Chen, and Le Song. Graph condensation via receptive field distribution matching. *arXiv preprint arXiv:2206.13697*, 2022.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *ICLR*, 2021a.

Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *NeurIPS*, 34:5186–5198, 2021b.

Xuran Pan, Xuan Jin, Yuan He, Shiji Song, Gao Huang, et al. Budgeted training for vision transformer. In *ICLR*, 2022.

Saro Passaro and C Lawrence Zitnick. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In *ICML*, pp. 27420–27438. PMLR, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *NeurIPS Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=n6SCkn2QaG`.

Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K Dokania, Philip HS Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual learning: What does matter? In *CVPR*, pp. 3698–3707, 2023.

Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1):779, 2022.

Nima Shoghi, Adeesh Kolluru, John R Kitchin, Zachary Ward Ulissi, C Lawrence Zitnick, and Brandon M Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. In *ICLR*, 2023.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian Roitberg. Outsmarting quantum chemistry through transfer learning. 2018.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10 (1):2903, 2019.

Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134, 2020.

Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J Medford, and David S Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture, 2024.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2022a.

Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *NeurIPS*, 35:9813–9827, 2022b.

A More Epochs or More Data?

To extend the findings presented in the main paper, we explore the trade-off between increasing the number of training epochs and expanding the dataset size under a fixed computational budget. Specifically, we aim to answer the following question:

💡 ***Given a fixed computational budget, is it more effective to train on a smaller dataset for more epochs or to train on a larger dataset for fewer epochs?***

Setup. To investigate this question, we compare two scenarios under the same computational budget of 10M samples: (1) training on 2M samples for 5 epochs, and (2) training on 1M samples for 10 epochs. We evaluate the performance of models pretrained on ANI-1x, Transition-1x, OC20, and OC22, and fine-tune them on the downstream datasets: rMD17, MD22, SPICE, and QM9. For comparison, we also include the results of JMP-L and JMP-S, which use 120M samples for 2 epochs.

Results. Table 4 presents the downstream performance for the two scenarios. Across all datasets, ANI-1x consistently achieves the best performance, regardless of whether the model is trained on 2M samples for 5 epochs or 1M samples for 10 epochs. For example, on rMD17, ANI-1x achieves a test error of 5.4 in both scenarios, outperforming JMP-S (6.7) and JMP-L (5.1). Similarly, on SPICE, ANI-1x achieves a test error of 5.08 (2M samples, 5 epochs) and 5.04 (1M samples, 10 epochs), compared to 5.71 for JMP-S and 4.75 for JMP-L.

Interestingly, increasing the number of epochs from 5 to 10 while reducing the dataset size from 2M to 1M does not significantly degrade performance for ANI-1x. This suggests that for highly aligned datasets like ANI-1x, training on fewer samples for more epochs can be as effective as training on more samples for fewer epochs. In contrast, for less aligned datasets such as OC20 and OC22, increasing the number of epochs only partially compensates for the reduced dataset size, as some tasks show similar performance while others experience noticeable degradation.

Takeaway. Our findings indicate that the trade-off between more epochs and more data depends on the alignment of the pretraining dataset with the downstream task. For highly aligned datasets like ANI-1x, training on fewer samples for more epochs can yield comparable performance. In contrast, for less aligned datasets, increasing the dataset size tends to be more beneficial. These results further show the importance of dataset quality and alignment, as quantified by CSI, in determining an effective pretraining strategy.

Table 4: Trade-off between increasing the number of samples and the number of epochs. We report the MAE for various downstream tasks while varying the pretraining sample count and epoch count simultaneously. $\mathcal{C}$, $\mathcal{N}$, and $\mathcal{E}$ denote the computational budget, number of samples, and number of epochs, respectively.

$\mathcal{C}$	$\mathcal{N}$	$\mathcal{E}$	**Upstream Data**	**Backbone**	**rMD17** (meV/Å)	**MD22** (meV/Å)	**SPICE** (meV/Å)	**QM9** (meV)
10M	2M	5	ANI-1x	GemNet-OC-S	**5.4**	**2.90**	**5.13**	**2.9**
			Transition-1x		10.1	3.73	7.55	3.2
			OC20		14.6	4.53	8.74	4.8
			OC22		16.0	5.20	10.73	5.7
10M	1M	10	ANI1x	GemNet-OC-S	**5.4**	**2.88**	**5.04**	**2.9**
			Transition1x		10.6	3.79	7.50	3.1
			OC20		14.8	4.67	10.16	4.9
			OC22		17.3	5.24	11.06	5.4

B Implementation Details

For both pretraining and fine-tuning experiments, we primarily follow the JMP hyperparameters. However, due to resource constraints requiring smaller batch sizes compared to JMP, we adjusted the learning rate to ensure training stability, as detailed below.

For pretraining, we use a batch size of 20 and a learning rate (LR) of 1e-4 for the small backbone (GemNet-OC-S). For the large backbone (GemNet-OC-L), the batch size is reduced to 12 to fit GPU memory. Additionally, when training with the OC22 dataset on the large backbone, a LR of 1e-4 caused gradient instability, thus we used a LR of 1e-5 for that particular run. Unless otherwise specified, each experiment is run for five epochs on the specified number of samples for each section of the paper. The best checkpoint is selected based on the performance in the validation set. To handle the large size of the upstream validation sets, validation is performed on a smaller subset of 2,000 samples.

For finetuning, we use the batch size specified in the JMP codebase and a default learning rate (LR) of 8e-5, except for cases where adjustments were needed to stabilize training. Specifically, we use 5e-5 for QMOF, 8e-4 for MatBench when pretrained on Transition1x.