Supplementary Information

A Implementation Details

A.1 Graph Neural Networks

Our implementation of CGCNN and TorchMD-Net utilizes the MatDeepLearn package [36] as a backbone. Graph representations are calculated through the algorithms released by the Open Catalyst Project (OCP) [38].

We report the main hyperparameters for CGCNN and TorchMD-Net below.

Table A.1: CGCNN hyperparameters for pre-training on MP Forces

Parameter	Value or description
Dimension 1	100
Dimension 2	150
Pre-graph convolution FC layers	1
Graph convolution	4
Post-graph convolution FC layers	3
Activation function	silu
Dropout rate	0
Pooling	Global add pool
Pool order	Early

Table A.2: TorchMD-Net hyperparameters for pre-training on MP Forces

Parameter	Value or description
Hidden channels	128
Number of filters	128
Number of layers	8
Number of RBF	50
RBF type	expnorm
Trainable RBF	True
Activation	silu
Attention activation	silu
Number of heads	8
Distance influence	both
Number of post layers	2
Post hidden channels	128
Pooling	Global add pool
Pool order	Early

A.2 Training Settings

Pre-training datasets, namely MP Forces and MP Forces Relaxed, are split on a `train:test:val` ratio of `0.8:0.15:0.05`. All finetuning datasets are split on a `train:test:val` ratio of `0.6:0.2:0.2` to ensure consistency and fair comparison. Models are pre-trained for 100 epochs and finetuned for 200 epochs for all datasets.

For graph representation generation, we used a cutoff radius of 8.0 and a maximum neighbor count of 250.

In the case of pre-training via denoising, the amount of noise added to atomic positions is generated according to a normal distribution of zero mean and a standard deviation of 0.1.

Several different GPU models are used in this work, including NVIDIA A100 80GB, NVIDIA A100 40GB, and

NVIDIA A40 48GB. Pre-training of CGCNN and TorchMD-Net on MP Forces takes approximately 10 and 60 hours respectively on a single A100 80GB GPU.

B Errors on Main Results

Here we present tables detailing the standard errors associated with the main results. Each entry in the main results tables represents the average of 5 runs. To quantify the uncertainty, we calculate the error as the standard deviation divided by the square root of 5:

$$\text{Standard Error} = \frac{\text{Standard Deviation}}{\sqrt{\#\text{ of Measurements}}}$$

Table B.1: Standard errors associated with Table 2.

		JDFT	Phonons	Dielectric	GVRH	KVRH	Perovskites	2D	MOF	Surface	MP gap	MP form
	Baseline	±4.37	±4.18	±0.0169	±0.00124	±0.00118	±0.000294	±0.00366	±0.0104	±0.00121	±0.00351	±0.000516
Forces	1:0:0	±5.46	±3.86	±0.0535	±0.00173	±0.00058	±0.000741	±0.00326	±0.0029	±0.000717	±0.00198	±0.00216
	0:1:0	±3.32	±4.40	±0.00634	±0.00524	±0.0178	±0.000529	±0.00546	±0.00229	±0.000964	±0.00545	±0.00296
	0:1:1	±5.82	±5.40	±0.0277	±0.00146	±0.00180	±0.000709	±0.00569	±0.00339	±0.00137	±0.00258	±0.00181
	1:1:1	±5.14	±2.21	±0.0129	±0.00246	±0.000986	±0.000511	±0.00481	±0.00101	±0.00110	±0.00262	±0.00143
	1:500:500	±3.49	±3.20	±0.0265	±0.000812	±0.00126	±0.000298	±0.00546	±0.00235	±0.000662	±0.00180	±0.00336
	Derivative-based denoising	±2.17	±7.74	±0.0343	±0.00207	±0.00109	±0.000422	±0.00616	±0.0513	±0.00209	±0.00365	±0.000975
	Prediction head denoising	±3.87	±2.39	±0.0360	±0.00212	±0.0132	±0.000628	±0.00585	±0.00483	±0.000874	±0.00268	±0.00350

Table B.2: Standard errors associated with Table 3.

		JDFT	Phonons	Dielectric	GVRH	KVRH	Perovskites	2D	MOF	Surface	MP gap	MP form
	Baseline	±2.32	±4.34	±0.0356	±0.00132	±0.0021	±0.0000842	±0.00421	±0.00138	±0.00117	±0.00181	±0.000339
Forces	1:0:0	±5.46	±3.86	±0.0535	±0.00173	±0.00058	±0.000741	±0.00326	±0.0029	±0.000717	±0.00198	±0.00216
	0:1:0	±3.32	±4.40	±0.00634	±0.00524	±0.0178	±0.000529	±0.00546	±0.00229	±0.000964	±0.00545	±0.00296
	0:1:1	±5.82	±5.40	±0.0277	±0.00146	±0.00180	±0.000709	±0.00569	±0.00339	±0.00137	±0.00258	±0.00181
	1:1:1	±5.14	±2.21	±0.0129	±0.00246	±0.000986	±0.000511	±0.00481	±0.00101	±0.00110	±0.00262	±0.00143
	1:500:500	±3.49	±3.20	±0.0265	±0.000812	±0.00126	±0.000298	±0.00546	±0.00235	±0.000662	±0.00180	±0.00336
	Derivative-based denoising	±3.25	±7.66	±0.0333	±0.00122	±0.000833	±0.000259	±0.00831	±0.00223	±0.00086	±0.00207	±0.000333
	Prediction head denoising	±3.66	±12.7	±0.0206	±0.00198	±0.00052	±0.000558	±0.00761	±0.0039	±0.000587	±0.00109	±0.000261

C Parity Plots of Forces For Pre-trained Models

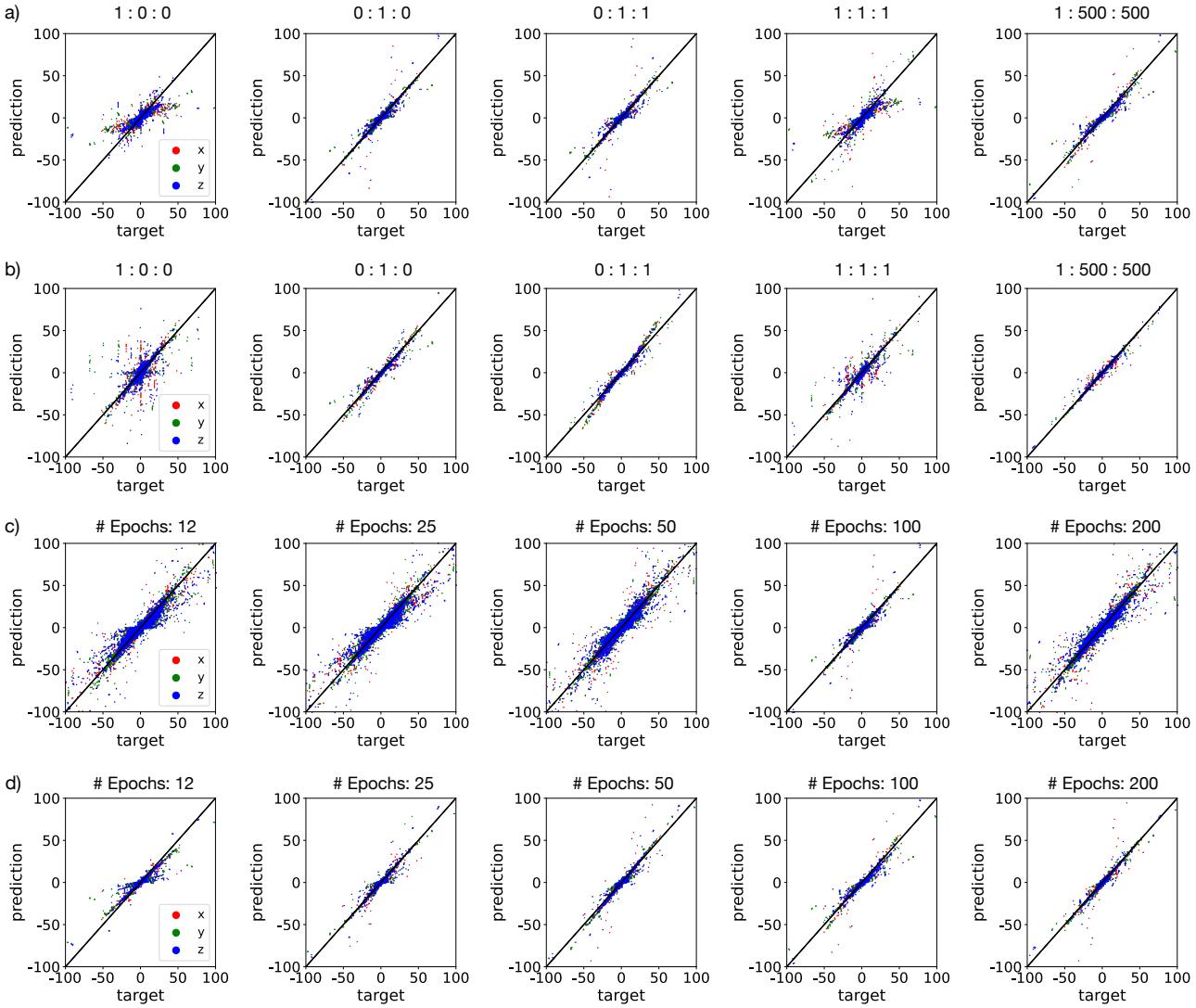

Figure C.1: Parity plots of forces prediction versus ground-truth labels. Axes are trimmed to $[-100, +100]$. a) CGCNN pre-trained models. b) TorchMD-Net pre-trained models. c) CGCNN pre-trained for different number of epochs on ratio $\lambda_{\text{energy}} : \lambda_{\text{forces}} : \lambda_{\text{stress}} = \texttt{0:1:0}$. d) CGCNN pre-trained for different number of epochs on ratio $\lambda_{\text{energy}} : \lambda_{\text{forces}} : \lambda_{\text{stress}} = \texttt{1:500:500}$.

D Embedding Visualization

To understand the nature of the benefit to finetuning performance from pre-training with energies and forces, we visualized the node-level embeddings for the CGCNN and TorchMD-Net models trained under the different ratios. Specifically, principal component analysis (PCA) is applied to 50,000 uniformly selected atom embeddings from the training split in the MP Forces dataset for a visualization in the two-dimensional space. It should be noted that both the derivative-based denoising and prediction head denoising models are trained on the smaller MP Forces Relaxed dataset. Consequently, it is possible that these two denoising models have not been exposed to some of the 50,000 atoms in the dataset.

We observe that the embeddings for CGCNN prediction head denoising is dissimilar from the rest. Our hypothesis is that this divergence stems from the fact that this particular model hasn't effectively learnt during the pre-training phase, as supported by its training loss barely decreasing and plateauing very early on.

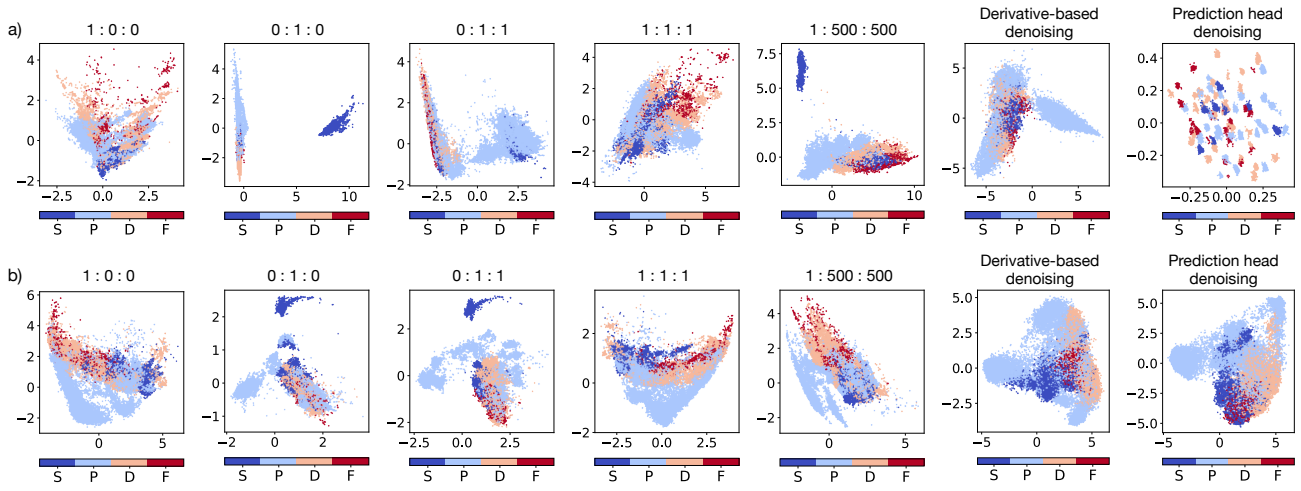

Figure D.1: Visualization of atom embedding space wherein atoms are categorized into S, P, D and F blocks on the periodic table. Principal component analysis (PCA) is applied to all atomic representations in the training split of the MP Forces dataset, and 50,000 uniformly selected atoms are chosen to be plotted. a) CGCNN pre-trained models. b) TorchMD-Net pre-trained models.