SELF-CONDITIONED DENOISING FOR ATOMISTIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of large-scale pre-training in language and vision has inspired the development of foundation models for the physical sciences; however, label-free pre-training methods for atomistic data remain under-investigated. To date, supervised pre-training on DFT force-energy labels has provided the largest benefit to downstream property prediction, outperforming self-supervised learning (SSL) methods that remain limited to ground-state geometries or single domains of atomistic data. We address these shortcomings with Self-Conditioned Denoising (SCD), a backbone-agnostic reconstruction objective in which the model's own embedding of a clean geometry, its self-embedding, conditions the denoising of a corrupted copy. SCD applies to any atomistic domain, including small molecules, proteins, periodic materials, and non-equilibrium geometries. With backbone and pre-training data matched, SCD reduces error by 20–45% relative to standard node denoising on QM9 targets, outperforming prior denoising variants on most targets, and matches the performance of supervised force-energy pre-training on identical geometries. A small, fast GNN pre-trained by SCD is competitive with or superior to larger models pre-trained on far larger labeled or unlabeled datasets across multiple domains. We also fine-tune leading foundation potentials (MACE-OFF, Orb-v3, UMA-S) for property prediction; SCD models remain competitive with them at one to two orders of magnitude less pre-training compute.

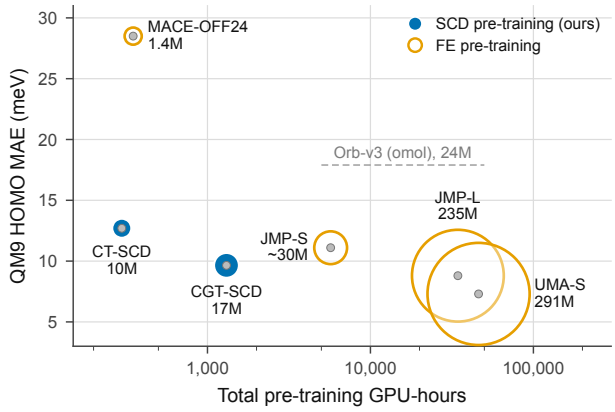

Figure 1: **Pre-training compute vs. QM9 HOMO error.** Marker area is proportional to parameter count. Filled markers: our SCD pre-trained models (CT and CGT); rings: force–energy pre-trained models (JMP-S/L; MACE-OFF24 and UMA-S from Appendix A). The dashed segment marks Orb-v3 (omol25 pretrained), whose pre-training hours are unreported (estimated range 5,000–50,000 GPU-hours). SCD hours were measured on NVIDIA A100 GPUs at fp32; JMP, MACE-OFF24 and UMA-S hours are as reported or estimated by their authors on different hardware (Table 6). The ET baseline (20.3 meV) is omitted. Results for all ten QM9 targets are in Table 12.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Foundation models have transformed the development pipeline for machine learning in vision and language. The prevailing paradigm pre-trains on large unlabeled datasets for generalizable representations, before subsequent training on smaller, task-specific, labeled data. This success has motivated analogous foundation models in the atomistic sciences, spanning chemistry, structural biology, and materials science (Yuan et al., 2025; Zhang et al., 2025), with the promise of accelerated simulation, structure-to-property prediction, and inverse design.

In chemistry and materials science, the most notable progress toward this vision has come from machine-learned interatomic potentials (MLIPs). Pre-trained MLIPs such as UMA (Wood et al., 2025), MACE (Batatia et al., 2025) and Orb (Rhodes et al., 2025) reach near-DFT accuracy in force-energy prediction through supervised pre-training on large, diverse atomistic datasets. Although often termed 'foundation models' (or more appropriately 'foundation potentials'), they differ from their language and vision counterparts in that they are pre-trained exclusively by supervised learning on precomputed forces and energies. Scaling this recipe is expensive; generating forces and energies for the 100M geometries in OMol25 required 6.6 billion CPU core-hours (Levine et al., 2025).

In other fields, foundation models have depended on self-supervised learning (SSL) objectives that leverage vast, unlabeled datasets. However, unlike images and text, which are often collected without labels, many atomistic datasets are computationally derived (e.g., via DFT) and inherently include force-energy labels. Although force and energy values vary depending on the DFT functional used, prior work has overcome this by using multiple prediction heads and demonstrating clear benefits to supervised pre-training on large, diverse datasets containing labels derived from multiple functionals (Wood et al., 2025; Shoghi et al., 2023). Consequently, supervised pre-training with DFT labels remains the dominant paradigm, outperforming existing SSL methods on simulation and property prediction benchmarks (Shoghi et al., 2023; Zaidi et al., 2022). However, the fraction of atomistic data carrying true DFT labels is likely to shrink as more geometries are produced by MLIP-driven simulation, relaxation, and generative models (Wohlwend et al., 2025; Passaro et al., 2025; Lemos et al., 2025; Wood et al., 2025; Zeni et al., 2023; Batatia et al., 2025; Li et al., 2024b; Morehead et al., 2026). Nevertheless, adapting SSL methods from other fields has remained challenging due to the unique attributes of 3D atomistic data. Atomistic data has no fixed size (samples ranging from 3 to over 3000 atoms), samples may have periodic boundaries, and model internal representation must be sensitive to small changes in atomic position to achieve accurate predictions in geometrically sensitive tasks (structural relaxation, simulation, binding affinity prediction, etc.).

Prior work in computer vision suggests that self-supervised pre-training can yield more interpretable and generalizable representations than supervised learning (Chen et al., 2021; Oquab et al., 2024; Siméoni et al., 2025), achieving performance competitive with supervised models even on identical datasets (He et al., 2021; Tomasev et al., 2022). This is because supervised learning only requires a model to represent features of the training data that are relevant to a specific task, becoming invariant to uncorrelated features. In contrast, generative SSL encourages a model to represent the full feature distribution within a training dataset, often producing internal representations with superior transferability. In this work, we demonstrate that this principle holds for 3D atomistic data as well. We do so with Self-Conditioned Denoising (SCD), in which the model's own embedding of a clean geometry conditions the denoising of a corrupted copy of that geometry. Along the way, we provide a systematic evaluation of fine-tuned foundation potentials as property predictors (Appendix A), a comparison that has received little attention.

2 BACKGROUND

2.1 NODE DENOISING

Early atomistic SSL maximized mutual information between 3D geometries and lower-dimensional views such as 2D graphs or SMILES (Liu et al., 2022; Stärk et al., 2022; Hu et al., 2020). These methods improved predictions from 2D inputs but underperformed supervised 3D models and fail to embed conformational state information necessary for tasks such as force/energy prediction or docking.

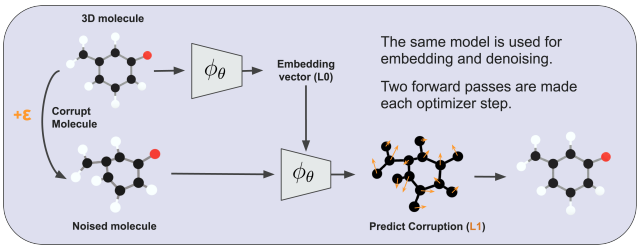

Figure 2: **Self-Conditioned Denoising pre-training.** A first forward pass embeds the clean geometry into a single invariant (L0) embedding vector. This self-embedding conditions a second forward pass (hence self-conditioned), in which the model predicts the per-atom corruption noise ($\boldsymbol{\varepsilon}$) from its equivariant (L1) output. The same model is used for both passes, and both passes occur in each optimizer step.

Node denoising emerged as a more general alternative that operates directly on 3D point clouds. Given a structure with N atoms $\mathbf{x} \in \mathbb{R}^{N\times 3}$, we sample Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ to form a corrupted geometry $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\varepsilon}$, and train a network ϕ_θ to predict the noise on each atom:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{x},\boldsymbol{\varepsilon}} \left[\|\phi_\theta(\tilde{\mathbf{x}}) - \boldsymbol{\varepsilon}\|_2^2 \right] \tag{1}$$

Unlike contrastive objectives that promote multi-view invariance, noise prediction encourages equivariance to geometric perturbations, making it effective for learning geometric features of local atomic environments. Node denoising has proven useful both as an auxiliary loss (Godwin et al., 2022; Liao & Smidt, 2022; Liao et al., 2023) and as a standalone pre-training task (Zaidi et al., 2022; Liu et al., 2023; Fang et al., 2025; Jiao et al., 2025; Yan et al., 2025). In computer vision, denoising is likewise an effective self-supervised objective in its own right (Xiang et al., 2023), and conditioning the denoiser on a compact latent of the clean input, learned end to end, can produce transferable representation (Preechakul et al., 2022; Hudson et al., 2024). SCD brings this bottleneck-conditioned design to atomistic data.

2.2 State of the Art Pre-Training

Despite its generality, standard node denoising has struggled to match supervised force-energy (FE) pre-training, largely because it fails to benefit from non-ground-state ('non-equilibrium') geometries in property prediction (Feng et al., 2023). Early interpretations suggested that denoising approximates learning a 'near-equilibrium' force field (Zaidi et al., 2022), and Wang et al. (2023b) report gains from denoising non-equilibrium conformers when fine-tuning neural potentials. For high-energy configurations, however, the noise vectors $\boldsymbol{\varepsilon}$ (pointing from corrupted to uncorrupted positions) are unlikely to resemble true forces, and the difference between a corrupted input and a valid higher-energy configuration becomes ambiguous, so denoising non-equilibrium geometries is ill-posed without additional information. One solution is a force encoding (Liao et al., 2024), but this requires force-energy labels and is no longer SSL.

Others have shown little correlation between noise vectors and true DFT forces even for ground-state geometries (Ni et al., 2023); Appendix B gives our reading of why denoising helps. Variations on node denoising have improved performance at the cost of domain-specific constraints, such as torsional noise in *Frad* (Ni et al., 2024) or pseudo-force targets in *SliDe* (Ni et al., 2023); others use specialized masking, block-wise denoising, hybrid contrastive objectives, or scaling to 200M geometries (Fang et al., 2025; Feng et al., 2024; Jiao et al., 2025; Zhou et al., 2023), but often sacrifice generality and remain limited to ground-state small molecules. To date, general-purpose SSL objectives have not matched supervised FE pre-training on downstream property prediction (Shoghi et al., 2023; Kim et al., 2025).

3 METHOD

3.1 THE LIMITATIONS OF NODE DENOISING

We identify three critical limitations of standard node denoising as a pre-training objective. **(s1) Locality.** GNNs often struggle with global representations and long-range correlations, and node denoising fails to produce strong global semantic learning: Gaussian noise on atom positions lacks long-range correlation across space, and small perturbations can be estimated accurately from a small, local context window. **(s2) Scalar vs. vector embedding pressure.** Equivariant GNNs partition node embeddings into channels by rotational symmetry group ($L = 0, 1, 2, \dots$). Noise prediction is an equivariant vector ($L = 1$; hereafter L1) objective and primarily impacts L1 channels; although gradients reach all parameters, there is little pressure on the invariant scalar embeddings ($L = 0$; hereafter L0) to develop good semantic representations, yet downstream property prediction is often roto-translation invariant and relies on the L0 channel. **(s3) Ambiguity with non-equilibrium structures.** As discussed in Section 2.2, without additional context a model cannot distinguish a corrupted input from a valid high-energy geometry, so standard node denoising does not benefit from non-equilibrium data.

3.2 SELF-CONDITIONED DENOISING

Self-Conditioned Denoising (SCD)[1] addresses all three shortcomings by adding a conditional embedding to the denoising objective of Equation 1 (Equation 2):

$$\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\|\phi_\theta(\tilde{\mathbf{x}}|\mathbf{c}) - \boldsymbol{\varepsilon}\|_2^2\right], \quad \mathbf{c} = f_\eta(\mathbf{x}) \tag{2}$$

Here, $\mathbf{c} \in \mathbb{R}^d$ is an embedding vector that represents the uncorrupted target geometry. This conditioning gives the model enough information to distinguish a corrupted sample from a valid target conformation (addressing s3). The conditional embedding itself then becomes a training objective that develops a meaningful invariant (L0) representation (addressing s2), which in turn pushes the embedding model f_η toward a better global, semantic representation (addressing s1). In SCD, the same model serves as both the embedding model f_η and the noise predictor ϕ_θ in each optimizer step (Equation 3):

$$\begin{aligned} &\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\|\phi_\theta(\tilde{\mathbf{x}}|\mathbf{c}) - \boldsymbol{\varepsilon}\|_2^2\right] = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\|\phi_\theta(\tilde{\mathbf{x}}|\phi_\theta(\mathbf{x})_{L0})_{L1} - \boldsymbol{\varepsilon}\|_2^2\right], \\ &\mathbf{c} = \phi_\theta(\mathbf{x})_{L0} \in \mathbb{R}^d, \qquad \phi_\theta(\tilde{\mathbf{x}}|\mathbf{c})_{L1} \in \mathbb{R}^{N\times 3} \end{aligned} \tag{3}$$

Algorithm 1 and Figure 2 summarize the implementation. First a forward pass is made on the uncorrupted geometry $\mathbf{x}$. The resulting invariant (L0) embeddings are sum-pooled across nodes and passed through a two-layer MLP embedding head, forming a single embedding per geometry. This information bottleneck encourages a semantic representation: one organized by chemical identity (composition, bonding, and conformation) rather than by extensive quantities such as atom count. The embedding then conditions a second forward pass on the corrupted geometry $\tilde{\mathbf{x}}$, whose L1 equivariant output predicts the per-node noise. The two noise scales in Algorithm 1 play distinct roles: σ_{corr} generates $\tilde{\mathbf{x}}$, while a much smaller regularization noise σ_{reg} is applied to the clean input of the embedding pass to reduce overfitting (Appendix B.1). To preserve unconditional behavior, we drop the conditional embedding for a random 20% of samples in each batch. After pre-training, downstream fine-tuning and inference require only a single pass without conditioning.

The embedding pass is trained end to end: gradients from the denoising loss flow through $\mathbf{c}$ into the first pass, so the embedding learns whatever representation best supports denoising. In principle, a d-dimensional embedding of a small molecule could encode its clean geometry outright and make denoising trivial; however, we do not observe this, because strong regularization (weight decay, σ_{reg} on the embedding pass, and condition dropout) penalizes memorization. Further, because $\mathbf{c}$ is invariant to rotation, translation, and atom ordering, reproducing exact positions and elements from it would require learning a canonical atom ordering – a harder task than denoising itself.

[1] Distinct from 'self-conditioning' in diffusion sampling (Chen et al., 2023), which conditions a denoiser on its own previous estimate of the clean sample with no gradient through it; in SCD the condition is a learned embedding of the clean input, trained end to end.

Algorithm 1 SCD pre-training and fine-tuning pipeline

Require: Dataset $\mathcal{D}$, noise scales $\sigma_{\text{corr}}, \sigma_{\text{reg}}$, model ϕ_θ
Ensure: Trained parameters θ
Pre-training
for each minibatch $\mathbf{x} \sim \mathcal{D}$ **do**
  $(\mathbf{x}, \tilde{\mathbf{x}}, \boldsymbol{\varepsilon}) \leftarrow \textsc{Corrupt}(\mathbf{x}, \sigma_{\text{corr}}, \sigma_{\text{reg}})$ ▷ create views
  $(\mathbf{c}_{L0}, \mathbf{v}_{L1}, \mathbf{y}_{L0}) \leftarrow \phi_\theta(\mathbf{x})$ ▷ self-embedding
  $(\tilde{\mathbf{c}}_{L0}, \tilde{\mathbf{v}}_{L1}, \tilde{\mathbf{y}}_{L0}) \leftarrow \phi_\theta(\tilde{\mathbf{x}}, \mathbf{c}_{L0})$ ▷ conditional denoising
  $\mathcal{L}_{\text{pre}} \leftarrow \text{MSE}(\tilde{\mathbf{v}}_{L1}, \boldsymbol{\varepsilon})$
  $\theta \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}_{\text{pre}}$
end for
Fine-tuning / inference
for each labeled sample $(\mathbf{x}, \mathbf{y})$ **do**
  $(\mathbf{c}_{L0}, \mathbf{v}_{L1}, \mathbf{y}_{L0}) \leftarrow \phi_\theta(\mathbf{x})$
  $\mathcal{L}_{\text{fine}} \leftarrow \text{loss}(\mathbf{y}_{L0}, \mathbf{y})$
  $\theta \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}_{\text{fine}}$
end for

Table 1: Summary of pre-training and benchmarking datasets.

Pre-training dataset	Domain	Unique structures	Total geometries	Origin / functional
PCQ (PCQM4Mv2)	Small molecules	3.378M	3.378M	B3LYP/6-31G* DFT
GEOM10	Small molecules	304,339	2.7M	GFN2-xTB, conformers
AMP20 (Alex-MP-20)	Periodic materials	675,204	675,204	PBEsol DFT
SAIR	Protein–ligand	888k	4.4M	Boltz-1x, conformers
ALL (AllAtoms)	Above, mixed	5.2M	11.3M	union of the four above
OMol25 (4M split)	Mixed	—	4M	MD, non-equilibrium
Benchmarking dataset	**Domain**	**Training geometries**	**Task Chosen**	
QM9	Small molecules	110k	molecular properties (DFT)	
Matbench (mp gap)	Periodic materials	85k	band gap (DFT)	
LBA	Protein–ligand	4k	binding affinity ($-\log K_d$)	

PCQ (Nakata & Shimazaki, 2017), GEOM (Axelrod & Gómez-Bombarelli, 2022), AMP20 (Jain et al., 2013; Schmidt et al., 2023; 2022), SAIR (Lemos et al., 2025), OMol25 (Levine et al., 2025), Boltz-1x (Wohlwend et al., 2025), QM9 (Ramakrishnan et al., 2014), Matbench (Dunn et al., 2020), LBA (Townshend et al., 2022)

3.3 Backbone Architectures

Following established baselines, we implement SCD using TorchMD-Net (also known as Equivariant Transformer or *ET*) as the primary backbone (Thölke & De Fabritiis, 2022; Ni et al., 2024; 2023; Feng et al., 2024; Zaidi et al., 2022). For a fair comparison, all models we train use an embedding dimension of 256, 8 layers, and 8 heads, consistent with previous SSL studies (Table 10).

The ET backbone is a lightweight, attention-based GNN without the tensor products and angular embeddings now common in equivariant MLIPs. *Frad* (Ni et al., 2024) and *SliDe* (Ni et al., 2023) require explicit angular information and use a modified "Geometric Equivariant Transformer" (GET); for a direct comparison we evaluate SCD on both ET and the more expressive GET.

To incorporate conditional inputs into these non-conditional backbones, we replace the pre-attention layer normalization with Adaptive Layer Normalization (*AdaNorm*), analogous to its use in Diffusion Transformers (Peebles & Xie, 2023) and in representation-conditioned generation (Li et al., 2024a;b). AdaNorm has a negligible impact on inference speed and memory. In contrast, the GET backbone is approximately $4\times$ slower and has twice as many parameters as ET. We denote the conditional versions of the two architectures **CT** (Conditional TorchMD-Net) and **CGT** (Conditional Geometric TorchMD-Net). Implementation details and speed comparisons are in Appendix C.

3.4 Datasets and Tasks

Atomistic model training has often been siloed to one domain at a time: small molecules, biomolecules, or periodic materials. Although recent work pre-trains on increasingly diverse data (Fang

Table 2: **QM9 test error at matched backbone and pre-training data.** MAE on the QM9 test set; all pre-trained models used PCQ. Baseline (ET from scratch) and Coord (node denoising on ET) are the values reported by Thölke & De Fabritiis (2022) and Zaidi et al. (2022); Frad and SliDe are from their publications. SCD (CT, CGT) is this work, one run per value. Bold: best per target; underline: second best.

Target	**Baseline**	**Coord**	**Frad**	**SliDe**	**SCD**	**SCD**
Backbone	ET	ET	GET	GET	CT	CGT
Pre-training data	none	PCQ	PCQ	PCQ	PCQ	PCQ
HOMO (meV)	20.3	17.7	15.2	13.6	<u>12.7</u>	**9.65**
LUMO (meV)	17.5	14.3	13.7	12.3	<u>11.5</u>	**9.05**
Gap (meV)	36.1	31.8	27.8	26.2	<u>24.5</u>	**19.7**
ZPVE (meV)	1.84	1.71	1.42	1.52	**1.18**	<u>1.22</u>
U_0 (meV)	6.15	6.57	5.33	4.28	**3.58**	<u>3.96</u>
U (meV)	6.38	6.11	5.62	4.29	**3.50**	<u>4.11</u>
H (meV)	6.16	6.45	5.55	4.26	**3.52**	<u>4.00</u>
G (meV)	7.62	6.91	6.19	<u>5.37</u>	**5.29**	**5.29**
α (a_0^3)	0.059	0.0517	<u>0.0374</u>	**0.0366**	0.0377	0.0383
C_v (cal/mol K)	0.026	<u>0.020</u>	<u>0.020</u>	**0.019**	0.021	**0.019**

et al., 2025; Joshi et al., 2025; Shoghi et al., 2023), multi-domain pre-training remains under-explored. We evaluate models pre-trained on a range of single-domain datasets with varying conformational diversity, and on a combined multi-domain aggregate (Table 1). All pre-trained models are then benchmarked on three representative domain specific tasks: 1) QM9 HOMO energy for small molecules (Ramakrishnan et al., 2014), 2) Matbench band gap for periodic materials (Dunn et al., 2020), and 3) ligand binding affinity (LBA) for bio-molecules (Townshend et al., 2022). Dataset descriptions are in Appendix E.

4 Results

4.1 SCD Outperforms Prior Denoising Objectives on Matched Data

Given the same backbone and pre-training dataset (PCQ), SCD outperforms prior methods of node denoising on QM9 property prediction (Table 2). Relative to standard node denoising, aka 'Coord', (Zaidi et al., 2022)), SCD on the CT backbone lowers error by 19.6–45.5% on all QM9 targets except heat capacity (C_v, -5%). Relative to the published Frad and SliDe results on the GET backbone, SCD on either architecture (CT or CGT) is better on most targets despite minor differences in fine-tuning protocol (Appendix E): CGT-SCD improves on SliDe by 20–29% on HOMO, LUMO, gap and ZPVE and by 1.5–7.5% on the four energies, and is within 5% on α and C_v. Table 17 in the Appendix compares against methods that use different backbones and pre-training datasets.

4.2 Labels Are Not Required: SCD vs. Force-Energy Pre-Training

To compare SCD directly with supervised force-energy (FE) pre-training, we pre-train two CT models on the identical 4M split of OMol25: one by FE prediction using the provided labels, and one by SCD using only the geometries. Figure 3 summarizes downstream results for each pre-training dataset (absolute errors in Table 11, Appendix D.2). On the OMol25 pair, SCD performs roughly as well (MP gap, -1.1%) or slightly better (QM9 HOMO, +3.9%; LBA, +0.8%) than FE pre-training, without a single label. Each value is a single training run, so we read these margins as parity. The FE baseline shown is the best of several FE pre-training and fine-tuning protocols (see Appendix D.4 for details).

4.3 SCD Benefits From Pre-Training on Any Atomistic Data Source

Using all available atomistic data for pre-training remains a bottleneck for atomistic foundation models: prior SSL work has been limited to ground-state small molecules (Ni et al., 2024; 2023; Feng et al., 2024), and force-energy pre-training across DFT functionals often requires dataset-specific heads or other specialized components (Shoghi et al., 2023; Wood et al., 2025). We find that SCD pre-training benefits from every dataset tested. Table 3 shows how datasets from different domains,

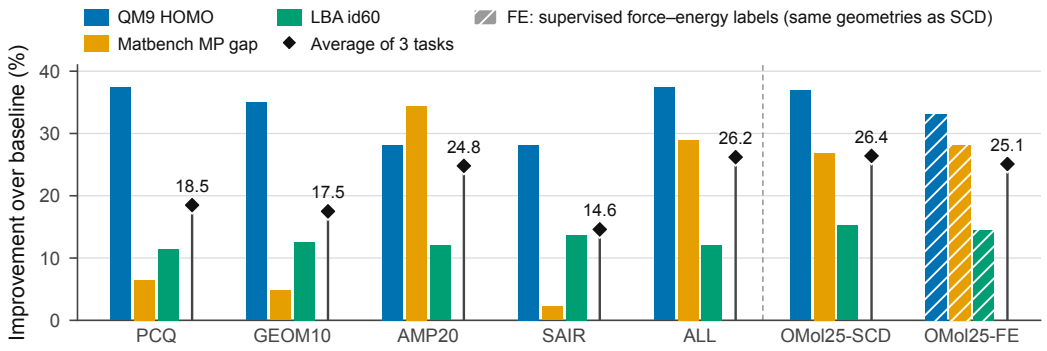

Figure 3: **Every pre-training source helps every task.** Improvement over the non-pre-trained CT baseline on QM9 HOMO (MAE), Matbench MP gap (MAE) and LBA id60 (RMSE) per SCD pre-training dataset; diamonds mark the three-task average. The two OMol25 models share identical geometries, without (SCD) and with (FE) force–energy labels. One run per model; absolute errors in Table 11.

Table 3: **SCD pre-training on diverse data sources.** QM9 HOMO error after SCD pre-training on ground-state and non-equilibrium geometries from varied sources; every configuration improves on the non-pre-trained baseline.

Pre-training Dataset	**Source / Quality**	**Unique Structures**	**Total Geom.**	**QM9 HOMO** (MAE, meV)
baseline				20.3
PCQ	DFT (B3LYP/6-31G*)	3,378,606	3,378,606	**12.7**
PCQ-300k	DFT (B3LYP/6-31G*)	300,000	300,000	13.0
GEOM10	xTB (Conformers)	304,339	2,791,929	13.2
GEOM1	xTB (Lowest Energy)	304,339	304,339	13.7
AMP20	DFT (PBEsol)	675,204	675,204	14.6
SAIR	Boltz-1x (Synthetic Conformers)	888,104	4.4M	14.6
ALL	Mixed Aggregate	5,246,253	11.3M	**12.7**
OMol25[1]	Non-equilibrium (MD)	—	3,986,754	12.8

[1] The number of unique molecules in the OMol25 4M split is not reported (Levine et al., 2025).

and of different conformational diversity, affect QM9 HOMO error. SCD benefits from including multiple conformers (GEOM1 vs GEOM10) and non-equilibrium structures (OMol25). We also find that large datasets are not always required for good results: pre-training on a random 10% subset of PCQ (300k samples) gives only slightly lower performance than the whole dataset (3.4M samples). Overall, the strongest predictor of downstream accuracy is the similarity between the pre-training and fine-tuning datasets. We also find that better-performing models have smoother, less extensive latent spaces; UMAP projections of the learned embeddings and their relation to downstream error are presented in Appendix D.1.

4.4 SCD Transfers Across Domains

SCD also transfers to materials and proteins. In Table 4, our 10M-parameter model pre-trained on 675k periodic materials reaches band gap accuracy within 2% of JMP-S, a ∼30M-parameter model trained on 120M force-energy labeled geometries (108M of them materials). Table 5 shows a similar outcome for ligand binding affinity (LBA) on protein-ligand complexes. Here we show how the SCD objective can exploit conditional relationships in atomistic data that are derived from domain knowledge. In the case of protein pocket-ligand pairs, we know that bound ligand conformations are determined by pocket geometry. With 'pocket-conditional ligand denoising' (CT-SCD-SAIR-Pocket) we can exploit this fact and provide a task-specific benefit to ligand binding affinity prediction, allowing our pre-trained models to outperform much larger models and achieve state of the art results

Table 4: **Matbench band gap prediction.** SCD models are competitive with far fewer parameters and unlabeled structures. Within each category, best is **bold**, second best underlined.

Model	MAE (eV)	Params	Data
No pre-training			
coGN	**0.156**	—	n/a
CT (Baseline)	0.186†	10M	n/a
JMP-S	0.235†	∼30M	n/a
JMP-L	0.228†	235M	n/a
MODNet	0.220	—	n/a
FE pre-trained			
JMP-L	**0.091**	235M	120M
JMP-S	0.121	∼30M	120M
CT-FE-OMol25	0.134†	10M	4M
HackNIP	0.150	$>$ 25M	$>$ 32M*
SCD pre-trained (ours)			
CT-SCD-AMP20	**0.123**	10M	675k
CT-SCD-ALL	0.132†	10M	11.3M
CT-SCD-OMol25	0.136†	10M	4M

coGN (Ruff et al., 2023); MODNet (De Breuck et al., 2021); JMP (Shoghi et al., 2023); HackNIP (Kim et al., 2025). † Fold 0; full results in Table 14. * Estimated from the Orb-v2 backbone.

Table 5: **Ligand binding affinity.** RMSE on the 30% (id30) and 60% (id60) sequence-identity splits. Best is **bold**, second best underlined, per split.

Model	RMSE ↓ id30 / id60	Params
No pre-training		
ProtNet	1.463 / 1.343	—
EPT-Scratch	1.378 / 1.277	30M
CT (Baseline)	1.510 / 1.386	10M
Other pre-trained		
Frad	1.365 / 1.213	14M
Uni-Mol	1.520 / 1.619	47.6M
EPT-Protein	1.329 / 1.235	30M
EPT-Multi	1.322 / 1.227	30M
ADiT-S	1.337 / 1.413	12M
ADiT-M	1.353 / 1.335	35M
ADiT-L	1.308 / 1.246	253M
SCD pre-trained (ours)		
CT-SCD-ALL	1.372 / 1.218	10M
CT-SCD-PCQ	1.332 / 1.226	10M
CT-SCD-SAIR	1.337 / 1.196	10M
CT-SCD-SAIR-Pocket	**1.304** / 1.200	10M
CT-SCD-OMol25	1.389 / **1.175**	10M

ProtNet (Wang et al., 2023a); EPT (Jiao et al., 2025); Frad (Ni et al., 2024); Uni-Mol (Zhou et al., 2023); ADiT (Shi et al., 2026). Full results in Table 13.

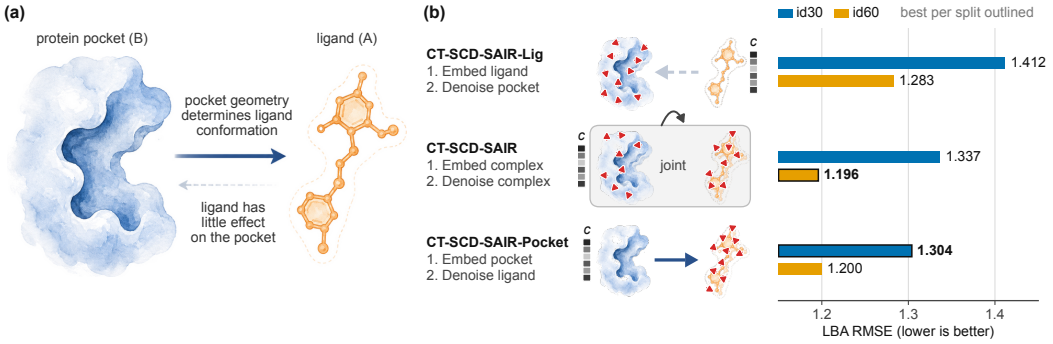

Figure 4: **Component-conditional denoising on pocket–ligand pairs. (a)** Ligand conformation depends strongly on pocket geometry (solid arrow); the reverse dependence is weak (dashed). B is the pocket and A the ligand in Equation 4. **(b)** Three SAIR pre-training objectives that differ only in the direction of conditioning; red triangles mark the denoised component and **c** the embedded one. LBA test RMSE on the id30 and id60 splits (Table 13) tracks the direction of conditioning; the best value per split is outlined.

(shown in Table 4.4). Whereas conditioning in the opposite direction ('ligand-conditional pocket denoising') harms performance as compared to joint denoising (Figure 4; Appendix D.3.1).

4.5 Compute Cost Against Larger Pre-Trained Models

Figure 1 places pre-training compute against QM9 HOMO error for SCD, force-energy pre-trained models, and the fine-tuned foundation potentials of Appendix A. A 10M-parameter CT pre-trained

by SCD for 298 GPU-hours (NVIDIA A100, fp32, no compilation) reaches lower error than EquiformerV2 trained from scratch and approaches JMP-S accuracy at one twentieth of its pre-training compute; including fine-tuning on ten QM9 targets, its total cost is also below training EquiformerV2 from scratch (Table 12, Appendix D.2), and it remains competitive with models pre-trained by supervised force-energy prediction on orders of magnitude more data (Shoghi et al., 2023). Hours for JMP, EquiformerV2 and the foundation potentials are as reported or estimated by their authors on different hardware, so the compute axis is indicative. Appendix A compares SCD with fine-tuned foundation potentials on all three tasks; full fine-tuning gave the best results in every case, and training only a head on a frozen potential did not reliably beat training the same architecture from scratch (Table 7). UMA-S reaches lower HOMO error than any SCD model, at roughly 150× the pre-training compute and 29× the parameters; MACE-OFF24 and Orb-v3 (trained on full OMol25) do not.

5 Discussion

SCD is a simple change to node denoising: embed the clean geometry, condition the denoiser on that embedding, and train both with one loss. Across eight pre-training datasets and three property-prediction domains, SCD improves every task, matches force-energy pre-training on identical geometries, and lets a 10M-parameter model compete with far larger supervised models at a fraction of the GPU-hours. Four observations follow.

All data is useful for pre-training. Every source tested (DFT ground states, xTB and diffusion-generated conformers, non-equilibrium MD) improved every task, and the benefit crosses domain boundaries: pre-training on periodic materials improves ligand binding affinity prediction by 12%, and OMol25, a non-periodic dataset spanning 83 elements, improves materials band gap predictions by 26.9% (Figure 3). What transfers is not periodicity but elemental coverage and conformational diversity, with part of the benefit likely from pre-smoothing of the representation space (Section 4.3).

Pre-training is not necessarily expensive. A simple, fast GNN pre-trained by SCD outperforms EquiformerV2 trained from scratch (Liao et al., 2023) in both accuracy and total compute when addressing multiple tasks (Figure 1; Table 12). A random 10% slice of PCQ (300k molecules) captures 96% of the full-dataset gain on QM9 HOMO, and 675k materials yield band gap accuracy within 2% of JMP-S (Shoghi et al., 2023), a 3× larger model pre-trained on 120M labeled geometries, 108M of them materials. Efficient pre-training was more cost effective than inductive bias through complex architectural elements.

Data relevance is key to performance. The strongest predictor of success was similarity between pre-training and fine-tuning data in elemental coverage and conformational diversity; for band gap prediction, inorganic pre-training data helped far more than organic-only data. CT-SCD-ALL had the best average across domains but did not surpass single-domain models on in-domain tasks; we attribute this to the limited expressivity of a 10M-parameter model rather than to representation conflicts, and expect larger models to gain more from multi-domain data.

Pre-training a foundation model does not require labels. When controlling for architecture and pre-training geometries (labeled vs unlabeled), SCD pre-training matches the performance of supervised force-energy pre-training in downstream property prediction tasks (Section 4.2). Further, SCD benefits from pre-training on every data source tested, including diffusion-generated conformers (SAIR). Conditioning is also a design space: embedding one component of a coupled system to denoise the other can exploit known physical relationships, and the direction matters (Figure 4; Appendix D.3.1). As MLIP-driven simulation and generative models make unlabeled geometries more plentiful than DFT labels, objectives like SCD offer a route towards utilizing generated data sources for representation learning.

Limitations. We use one backbone family (TorchMD-Net and its geometric variant) at 10M–17M parameters, three downstream tasks, and one training run per configuration, so the small margins in Section 4.2 are not statistically resolved. The Baseline and Coord columns of Table 2, and all Frad, SliDe, JMP, EquiformerV2, UniCorn and EPT numbers, are taken from their publications rather than re-run (Section 4.1). We have not ablated the self-conditioning against the architecture changes that accompany it (AdaNorm, drop-path, frozen element embeddings). How SCD scales with model size

and multi-domain data remains open, and on force prediction (MD17) SCD offers little advantage over prior methods (Appendix D.5).

Ethics statement

This work aims to improve self-supervised pre-training for machine learning in the atomistic sciences and may reduce the need for computationally expensive DFT labels during pre-training. The societal considerations that apply to atomistic foundation models in general apply here: such models and their derivative tools could lower the cost of discovery in many areas of science, including dual-use technologies. All datasets used are publicly available under their own licenses; the work involves no human subjects or personal data.

Reproducibility statement

The SCD objective is defined in Section 3.2 (Equations 2–3, Algorithm 1) and the backbone modifications in Section 3.3 and Appendix C. Pre-training and fine-tuning hyperparameters for every run are listed in Appendix F, dataset sources and preprocessing in Appendix E, the force-energy baselines in Appendix D.4, and the foundation-potential fine-tuning protocol in Appendix A.2. Every reported number is a single training run; Matbench values marked with a dagger are fold 0. Pre-training GPU-hours were measured on NVIDIA A100 GPUs at fp32 precision without compilation. Code and pre-trained weights will be made publicly available upon acceptance.

AI use statement

We did not use generative AI tools for any task on the required-disclosure list: generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing ingredients for or writing proofs, proposing or refining hypotheses, designing or giving feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic analysis, or interpreting results. We used generative AI tools (Claude models by Anthropic, accessed through Claude Code, 2026) for tasks on the recommended-disclosure list: restructuring and reformatting the LaTeX manuscript between venue templates, writing the plotting scripts for Figures 1, 3 and 4 from values transcribed from our result tables, editing text for readability, summarizing and searching related literature, checking the manuscript against the submission guidelines, and drafting this end matter. All AI-assisted text was reviewed by the authors against the results it describes, plotting scripts were checked against the source tables, and every citation added with AI assistance was verified by the authors against the cited paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation, 2022. URL https://arxiv.org/abs/2006.05531.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, John L. A. Gardner, Mikolaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Cheuk Hin Ho, Stephan Hofmann, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn,

Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2025. URL https://arxiv.org/abs/2401.00096.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2208.04202.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.02057.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, May 2017. doi: 10.1126/sciadv.1603015. URL https://doi.org/10.1126/sciadv.1603015.

Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Computational Materials*, 7(83), 2021. doi: 10.1038/s41524-021-00552-2.

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. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

Ada Fang, Zaixi Zhang, Andrew Zhou, and Marinka Zitnik. ATOMICA: Learning universal representations of intermolecular interactions. *bioRxiv*, 2025.

Rui Feng, Qi Zhu, Huan Tran, Binghong Chen, Aubrey Toland, Rampi Ramprasad, and Chao Zhang. May the force be with you: Unified force-centric pre-training for 3D molecular conformations, 2023. URL https://arxiv.org/abs/2308.14759.

Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniCorn: A unified contrastive learning approach for multi-view molecular representation learning. *arXiv preprint arXiv:2405.10343*, 2024.

Bowen Gao, Yinjun Jia, Yuanle Mo, Yuyan Ni, Weiying Ma, Zhiming Ma, and Yanyan Lan. ProFSA: Self-supervised pocket pretraining via protein fragment-surroundings alignment. *arXiv preprint arXiv:2310.07229*, 2023.

Jonathan Godwin, Michael Schaarschmidt, Alexander L Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple GNN regularisation for 3D molecular property prediction and beyond. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1wVvweK3oIb.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks, 2020. URL https://arxiv.org/abs/1905.12265.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. SODA: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.17901.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 2013.

Rui Jiao, Xiangzhe Kong, Li Zhang, Ziyang Yu, Fangyuan Ren, Wenjuan Tan, Wenbing Huang, and Yang Liu. An equivariant pretrained transformer for unified 3D molecular representation learning, 2025. URL https://arxiv.org/abs/2402.12714.

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

So Yeon Kim, Yang Jeong Park, and Ju Li. Leveraging neural network interatomic potentials for a foundation model of chemistry. *arXiv preprint arXiv:2506.18497*, 2025. URL https://arxiv.org/pdf/2506.18497.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025. doi: 10.1021/jacs.4c07099. URL https://doi.org/10.1021/jacs.4c07099. arXiv:2312.15211.

Pablo Lemos, Zane Beckwith, Sasaank Bandi, Maarten Van Damme, Jordan Crivelli-Decker, Benjamin J Shields, Thomas Merth, Punit K Jha, Nicola De Mitri, Tiffany J Callahan, et al. SAIR: Enabling deep learning for protein-ligand interactions with a synthetic structural dataset. *bioRxiv*, pp. 2025–06, 2025.

Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, et al. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://arxiv.org/abs/2312.03701.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. *arXiv preprint arXiv:2410.03655*, 2024b.

Yi Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022. doi: 10.48550/arXiv.2206.11990. URL https://arxiv.org/abs/2206.11990.

Yi Liao, Tess Smidt, Brandon Wood, and Abhishek Das. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023. doi: 10.48550/arXiv.2306.12059. URL https://arxiv.org/abs/2306.12059.

Yi-Lun Liao, Tess Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields, 2024. URL https://arxiv.org/abs/2403.09549.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry, 2022. URL https://arxiv.org/abs/2110.07728.

Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching, 2023. URL https://arxiv.org/abs/2206.13602.

Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Computational Materials*, 8(1):231, 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00921-5. URL https://doi.org/10.1038/s41524-022-00921-5.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Tynan Perez, Samuel Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, Hrushikesh Sahasrabuddhe, Pietro Liò, Tommi Jaakkola, Rafael Gómez-Bombarelli, Rex Ying, N. Benjamin Erichson, and Michael W. Mahoney. Zatom-1: Towards a multimodal foundation model for 3D molecules and materials, 2026. URL https://arxiv.org/abs/2602.22251. arXiv:2602.22251.

Maho Nakata and Tomomi Shimazaki. PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, 2017. ISSN 1549-9596. doi: 10.1021/acs.jcim.7b00083. URL https://doi.org/10.1021/acs.jcim.7b00083.

Yuyan Ni, Shikun Feng, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Sliced denoising: A physics-informed molecular pre-training method. *arXiv preprint arXiv:2311.02124*, 2023.

Yuyan Ni, Shikun Feng, Xin Hong, Yuancheng Sun, Wei-Ying Ma, Zhi-Ming Ma, Qiwei Ye, and Yanyan Lan. Pre-training with fractional denoising to enhance molecular property prediction. *Nature Machine Intelligence*, 2024. doi: 10.1038/s42256-024-00900-z. URL https://www.nature.com/articles/s42256-024-00900-z.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619–10629, 2022.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale, 2025. URL https://arxiv.org/abs/2504.06231.

Robin Ruff, Patrick Reiser, Jan Stühmer, and Pascal Friederich. Connectivity optimized nested graph networks for crystal structures, 2023. URL https://arxiv.org/abs/2302.14102.

Jonathan Schmidt, Hai-Chen Wang, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals. *Scientific Data*, 9(1):64, 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01177-w. URL https://doi.org/10.1038/s41597-022-01177-w.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35(22): 2210788, 2023. doi: https://doi.org/10.1002/adma.202210788. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202210788.

Liang Shi, Zuobai Zhang, Huiyu Cai, Santiago Miret, Zhi Yang, and Jian Tang. Towards all-atom foundation models for biomolecular binding affinity prediction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=o0Qfsq1fK8.

Nima Shoghi, Adeesh Kolluru, John R Kitchin, Zachary W Ulissi, C Lawrence Zitnick, and Brandon M Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv preprint arXiv:2310.16802*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34:25244–25255, 2021.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves GNNs for molecular property prediction, 2022. URL https://arxiv.org/abs/2110.04126.

Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant transformers for neural network based molecular potentials, 2022. URL https://arxiv.org/abs/2202.02541.

Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet?, 2022. URL https://arxiv.org/abs/2201.05119.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions, 2022. URL https://arxiv.org/abs/2012.04035.

Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3D graph networks, 2023a. URL https://arxiv.org/abs/2207.12600.

Yuyang Wang, Changwen Xu, Zijie Li, and Amir Barati Farimani. Denoise pretraining on nonequilibrium molecules for accurate and transferable neural potentials. *Journal of Chemical Theory and Computation*, 19(15):5077–5087, 2023b. doi: 10.1021/acs.jctc.3c00289.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pp. 2024–11, 2025.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms, 2025. URL https://arxiv.org/abs/2506.23971.

Fang Wu, Yu Tao, Dragomir Radev, and Jinbo Xu. When geometric deep learning meets pretrained protein language models. *bioRxiv*, pp. 2023–01, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.09769.

Shaoheng Yan, Zian Li, and Muhan Zhang. GeoRecon: Graph-level representation learning for 3d molecules via reconstruction-based pretraining, 2025. URL https://arxiv.org/abs/2506.13174.

Eric C. Y. Yuan, Yunsheng Liu, Junmin Chen, Peichen Zhong, Sanjeev Raja, Tobias Kreiman, Santiago Vargas, Wenbin Xu, Martin Head-Gordon, Chao Yang, Samuel M. Blau, Bingqing Cheng, Aditi Krishnapriyan, and Teresa Head-Gordon. Foundation models for atomistic simulation of chemistry and materials, 2025. URL https://arxiv.org/abs/2503.10538.

Sheheryar Zaidi. Response to issue #3 on *shehzaidi/pre-training-via-denoising*: Training stability for homo/lumo tasks. https://github.com/shehzaidi/pre-training-via-denoising/issues/3#issuecomment-1670261990, 2023. GitHub issue comment.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. MatterGen: A generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *Foundations and Trends® in Machine Learning*, 18(4):385–912, 2025.

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. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6K2RM6wVqKu`.

A Fine-Tuning Foundation Potentials for Property Prediction

A.1 Results

Pre-trained foundation potentials have demonstrated very strong performance on force and energy prediction tasks. However, there has been relatively little effort to evaluate foundation potentials on other property prediction tasks. Here we evaluate three training methods: fine-tuning with a frozen backbone (only a prediction head is trained), full fine-tuning (pre-trained weights, all parameters trained), and training from scratch (same architecture, random initialization). In Table 6 we present the results from fully fine-tuning several popular pre-trained foundation potentials on the same three property prediction tasks used to evaluate our SCD models. We find that larger models perform well, especially UMA-S, but are very costly to pre-train, and relatively expensive to fine-tune fully. The comparative efficacy of our low cost pre-trained models suggests there may be room to improve pre-training efficiency for foundation potentials. Table 7 presents the results of training the same architectures from scratch, and fine-tuning with a frozen backbone – a low cost alternative to full parameter tuning. Notably, we find that training a prediction head with a frozen pre-trained backbone does not always outperform training from scratch (without pre-trained weights). Fine-tuning all parameters provided the best results on all three tasks tested (see Table 7c).

A.2 Fine-Tuning Methods

In order to fine-tune foundation potentials for property prediction, we replace each model's force-energy output layers with a simple prediction head. The same prediction head architecture is used across all tasks and pre-trained models. The head applies two per-atom linear layers, a pooling operation, then two more linear layers, all of width 256 with SiLU activations in between, except before pooling. The last layer is zero-initialized, so the initial prediction is zero. Sum pooling is used for QM9 HOMO and mean pooling for MP gap and LBA. Per-atom embeddings are taken from the backbone: for MACE, the concatenated scalar blocks of both message-passing layers; for Orb-v3, the final node embeddings; and for UMA, the scalar component of the node embedding, giving feature widths of 256, 256, and 128, respectively. All fine-tuning runs use AdamW with a linear warmup (5% of training length), cosine decay to LR/100, gradient clipping at $\|g\|_2 \leq 10$, and an exponential moving average of weights (decay 0.999) for all tasks; this was not applied to the SCD models. Weight decay is 10^{-4} and 10^{-2} for full-network and head-only training, respectively. Target labels are normalized (zero mean and unit standard deviation) based on the training split, and loss is computed by mean squared error. We then select the checkpoint with the minimum validation error, stop early after 10% of the epoch budget without improvement, and evaluate the test split once at that checkpoint. Head architecture, weight decay, schedule, and epoch budget are fixed across models; learning rate is the only tuned hyperparameter, searched over the same four-point grid, except for the LBA seq-id-60 cells, which reuse the learning rate selected on seq-id-30. Total batch size is 128 for QM9 and MP gap, and 8 for LBA; UMA required four GPUs to reach a batch size of 128 at 32 samples each. These training parameters are summarized in Table 8.

B Practical Notes on Node Denoising

B.1 Interpreting Noise Scales

We find that the scale of noise applied to atomistic structures plays a significant role in its utility as a pre-training or regularization tool. Here we formalize two distinct regimes:

Regularizing noise utilizes a small scale with respect to bond length (single bonds range from $\sim$1.0 to 2.0Å), such as $\sigma \approx 0.005$Å, in order to reduce overfitting. At this scale Gaussian noise does not alter single-double bond distinguishability, and is often used as an auxiliary task (Godwin et al., 2022; Liao & Smidt, 2022; Liao et al., 2023). We find that regularizing noise improves downstream results when used in SCD pre-training (on non-corrupted inputs) and performance during some fine-tuning tasks, but not all.

Corruption noise used for pre-training is typically one order of magnitude larger ($\sigma \approx 0.04$Å), and often perturbs molecular geometries enough to disrupt single/double/triple bond identities ($\pm$0.1Å),

Table 6: **Property Prediction with Fine-tuned Foundation Potentials**. Test set MAE or RMSE on property prediction tasks from three different domains. Model checkpoints were fine-tuned on tasks matching their pre-training domain. Lower is better. **Bold** denotes best; underline denotes second best.

		Pre-training	Pre-train	QM9	MP gap	LBA (RMSE)	
Model (checkpoint)	**Params**	**geometries**	**GPU-h**	**HOMO (meV)**	**(eV)**	**id30**	**id60**
MACE-OFF24 (med) (Kovács et al., 2025)	1.4M	1.16M	≈350*	28.5	—	—	—
MACE-MPA-0 (med) (Batatia et al., 2025)	9.06M	≥12M	≈2,600*	—	0.156	1.463	1.207
Orb-v3 (omol) (Rhodes et al., 2025)	24M	≥100M	unpublished	17.9	—	1.348	1.314
Orb-v3 (inf-mpa) (Rhodes et al., 2025)	24M	≥12M	unpublished	—	0.106	—	—
UMA-S (1p2) (Wood et al., 2025)	291M	≥460M	≥46,080*	**7.3**	**0.097**	1.349	1.205
CT-SCD-PCQ	10M	3.4M	298	12.7	0.174	1.332	1.226
CT-SCD-OMol25	10M	4.0M	1,790	12.8	0.136	1.389	**1.175**
CT-SCD-AMP20	10M	675k	240	14.6	0.122	1.408	1.219
CT-SCD-SAIR-Pocket	10M	4.4M	390	—	—	**1.304**	1.200
CT-SCD-ALL	10M	5.2M	2,180	12.7	0.132	1.372	1.218
CGT-SCD-PCQ	17M	3.4M	1,308	9.65	—	—	—

*Pre-training GPU-h are estimated; these figures are not published by the model authors. Hardware differs between rows (A100 for MACE-OFF24, H200 for UMA-S), so the totals are indicative rather than directly comparable.

but without moving atoms far enough to be considered bond breaking. We find that noise scales much larger than this harm pre-training performance.

B.2 Reinterpreting Denoising for Atomistic Systems

Many previous works have suggested a relationship between denoising objectives and learning 'force fields'; however, we do not find this to be a useful hypothesis for why denoising benefits representation. It is often implied that the 'field' learned by noise prediction is beneficial due to its similarity to DFT forces, yet, as noted in Section 2.2, noise vectors on ground-state geometries correlate poorly with true DFT forces. Consistent with this, we find that for our models trained on ground-truth DFT forces, adding an auxiliary node denoising objective harms performance on property prediction, suggesting a low similarity between these targets (see Table 16).

Instead, we hypothesize that denoising objectives succeed because they encourage the development of smooth, distinguishable local atom embeddings. SCD complements this by introducing a global semantic objective. Through this lens, we attribute the transferability of force-energy pre-training to its demand for high geometric sensitivity and representational complexity, rather than to its explicit connection to physical principles. Force prediction encourages equivariant sensitivity to local perturbations, while energy prediction promotes global representation. Based on this hypothesis, we suggest that self-supervised learning methods should employ reconstruction or de-corruption objectives that challenge models with both local and global objectives similar to force-energy pre-training and SCD.

C Backbone Architecture and Implementation

Self-Conditioned Denoising pre-training can be implemented with any architecture designed for diffusion. To modify an existing non-conditional GNN for SCD pre-training, we replace the pre-norm of each message passing layer with an adaptive layer norm (AdaNorm), similar to a diffusion transformer (Peebles & Xie, 2023).

Figure 5 diagrams our implementation of AdaNorm used for SCD. Because SCD pre-training requires two forward passes per-step we find that including drop path with a drop probability of 0.1 improves training stability and provides an additional form of regularization. Our drop path is implemented to jointly drop both L0 and L1 embeddings. During pre-training we also freeze atom type/element embeddings. Without freezing element embeddings, we find that element embeddings become vanishingly small, leading to downstream instability. On GitHub (Zaidi, 2023), previous authors have noted challenges (gradient explosion) during the downstream fine-tuning of models trained by

Table 7: **Isolating pre-training benefits from model architecture**. Test set MAE or RMSE on property prediction tasks under different training methods. Here we present results of training each architecture from scratch (a), with a frozen pre-trained backbone (b), and compare both to full parameter fine-tuning of pre-trained models (c). In (a) and (b), **bold** denotes best and underline second best. Lower is better.

(a) **Trained from scratch**: architecture kept, weights randomized.

		QM9	MP gap	LBA (RMSE)	
Architecture	**Params**	**HOMO (meV)**	**(eV)**	**id30**	**id60**
MACE-OFF24 (med)	1.4M	31.6	—	—	—
MACE-MPA-0 (med)	9.06M	—	0.289	1.479	1.363
Orb-v3 (omol)	24M	24.3	—	**1.457**	1.552
Orb-v3 (inf-mpa)	24M	—	0.176	—	—
UMA-S (1p2)	291M	**19.5**	**0.169**	1.462	**1.298**

(b) **Frozen backbone**: published weights fixed, only the head trained.

		QM9	MP gap	LBA (RMSE)	
Model (checkpoint)	**Params**	**HOMO (meV)**	**(eV)**	**id30**	**id60**
MACE-OFF24 (med)	1.4M	41.1	—	—	—
MACE-MPA-0 (med)	9.06M	—	0.160	1.520	1.442
Orb-v3 (omol)	24M	50.4	—	**1.457**	1.403
Orb-v3 (inf-mpa)	24M	—	**0.117**	—	—
UMA-S (1p2)	291M	**35.0**	0.133	1.457	**1.302**

(c) **Comparing methods**: scratch / frozen / fine-tuned. **Bold** marks the best method for each model and task.

Model (checkpoint)	**HOMO (meV)**	**MP gap (eV)**	**id30**	**id60**
MACE-OFF24 (med)	31.6 / 41.1 / **28.5**	—	—	—
MACE-MPA-0 (med)	—	0.289 / 0.160 / **0.156**	1.479 / 1.520 / **1.463**	1.363 / 1.442 / **1.207**
Orb-v3 (omol)	24.3 / 50.4 / **17.9**	—	1.457 / 1.457 / **1.348**	1.552 / 1.403 / **1.314**
Orb-v3 (inf-mpa)	—	0.176 / 0.117 / **0.106**	—	—
UMA-S (1p2)	19.5 / 35.0 / **7.3**	0.169 / 0.133 / **0.097**	1.462 / 1.457 / **1.349**	1.298 / 1.302 / **1.205**

standard node-denoising pre-training (aka coord) – likely a result of small or zeroed out embedding vectors.

Table 9 shows a breakdown of parameter count, speed, and memory usage for the architectures used in this work. Each architecture shown uses the same embedding size (256), number of layers (8), and number of attention heads (8). CT and CGT correspond to the modified architectures of ET and GET that incorporate conditional layer norms. In this table, we can see that adding conditional guidance to an architecture has little effect on its speed and memory usage as compared to the addition of edge updates and angle embeddings. CT is overall 4 times faster, almost half the size, and has half the peak memory usage as GET.

Table 8: **Training hyperparameters for foundation-potential benchmarking.**

(a) Settings held constant across every run.

Setting	Value
Optimizer	AdamW, β_1=0.9, β_2=0.999, ϵ=10^{-8}
Learning-rate schedule	5% linear warmup, then cosine decay to LR/100
Weight decay	10^{-4} (from scratch, fine-tuned); 10^{-2} (frozen)
Weight averaging	Exponential moving average, decay 0.999
Head architecture	2 linear layers $\rightarrow$ pool $\rightarrow$ 2 linear layers; width 256, SiLU except before pooling
Head pooling	sum (QM9 HOMO); mean (MP gap, LBA)
Epoch budget (QM9 / MP gap / LBA)	400 / 250 / 150
Early-stop patience	10% of the budget (40 / 25 / 15)
Early-stop minimum change	0.2 meV / 1.0 meV / 0.01 RMSE
Total batch size (QM9 / MP gap / LBA)	128 / 128 / 8
Precision	fp32 (MACE); bf16 (Orb-v3, UMA)
Hardware	NVIDIA A100 (40 GB); UMA uses 4 cards at 32 samples each

(b) Learning rate selected for each training method, given as scratch / frozen / fine-tuned.

Task	MACE	Orb-v3	UMA-S
QM9 HOMO	$(1/1/0.3) \times 10^{-3}$	$(0.1/1/0.3) \times 10^{-3}$	$(1/1/0.1) \times 10^{-3}$
MP gap	$(1/1/0.3) \times 10^{-3}$	$(0.03/1/0.03) \times 10^{-3}$	$(1/1/0.1) \times 10^{-3}$
LBA	$(0.1/1/0.3) \times 10^{-3}$	$(0.1/3/0.1) \times 10^{-3}$	$(0.1/3/0.1) \times 10^{-3}$

Table 9: **Architecture Efficiency Comparison**. Benchmarked using QM9 molecules, with batch size 128, averaged over 32 batches, on a NVIDIA RTX A5000 GPU, at fp32 precision with `torch.compile` off. All models share identical embedding dimensions, layer counts, and attention heads. Parameter counts in this table encompass only backbone architecture, excluding the embedding and prediction heads. All metrics shown are computed from singular forward passes.

(a) **Absolute Metrics**: Latency (lower is better), Throughput (higher is better), and Memory.

Model	Params	Latency (ms) ↓		Throughput (samp/s) ↑		Avg Mem (GB)		Peak Mem (GB)	
	(M)	Eval	Train	Eval	Train	Eval	Train	Eval	Train
ET	6.5	65.7	69.1	1946	880	0.04	0.13	1.07	8.47
CT	9.2	70.6	74.5	1812	802	0.05	0.15	1.08	9.00
GET	13.4	264.9	369.5	483	194	0.07	0.18	1.62	20.37
CGT	16.0	262.8	379.3	487	189	0.08	0.20	1.63	20.72

(b) **Relative Metrics**: Normalized to ET baseline (1.0×).

Model	Param Ratio	Rel. Speed ↑		Rel. Avg Mem ↓	
		Eval	Train	Eval	Train
ET	1.00×	1.00×	1.00×	1.00×	1.00×
CT	1.40×	0.92×	0.93×	1.25×	1.15×
GET	2.00×	0.25×	0.19×	1.67×	1.38×
CGT	2.45×	0.25×	0.18×	1.90×	1.53×

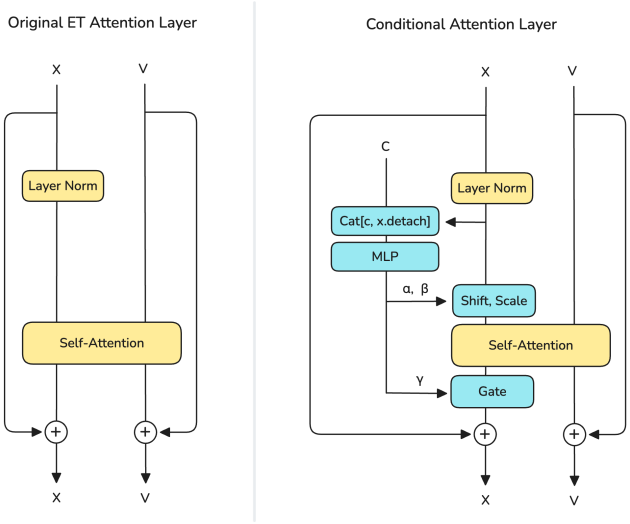

Figure 5: **Adaptive layer normalization for SCD:** On the left we illustrate a simplified diagram of the original residual layer block in TorchMD-Net (ET). Here, 'X' represents invariant (L0) embeddings while 'V' represents equivariant vector embeddings (L1), and 'C' represents the conditioning embedding. On the right, we show how we modify the original residual block with a conditional scale, shift, and gate of the invariant embeddings. In the gate block the gamma vector is attenuated by a tanh function and multiplied to the output of the ET Attention block. For this work, we use a two layer MLP for conditioning in each layer.

Table 10: Backbone architecture hyperparameters used in this work. CT-small is only used in the MD17 benchmark.

Hyperparameter	**CT**	**CGT**	**CT-small**
Layers	8	8	6
Attention heads	8	8	8
Embedding dimension	256	256	128
Layer pooling	sum	sum	sum
Graph cutoff (Å)	5	5	5
L0 head pooling	sum	sum	sum

D SUPPLEMENTARY RESULTS

D.1 BETTER MODELS HAVE SMOOTHER LATENT SPACES

Figure 6 provides UMAPs for QM9 molecule embeddings from both untrained and pre-trained models, colored by the number of atoms in each sample (N). We observe that an untrained ET model has a strong bias towards highly clustered, extensive representations – mean pooled atom embeddings that are highly correlated with the number of atoms in each molecule. We find that standard node denoising (coord) results in a smooth but highly extensive representation space (this was also observed by Li et al. (2024b)). We believe this is a consequence of features that are too biased towards local node geometry, rather than global semantic representation (non-extensive features). In models pre-trained by SCD, we observe greater smoothing and a reduced correlation with N, suggesting a more semantic representation. Empirically, we find models with a smoother, less extensive, embedding space reach a lower prediction error on QM9 targets (Section 4.3).

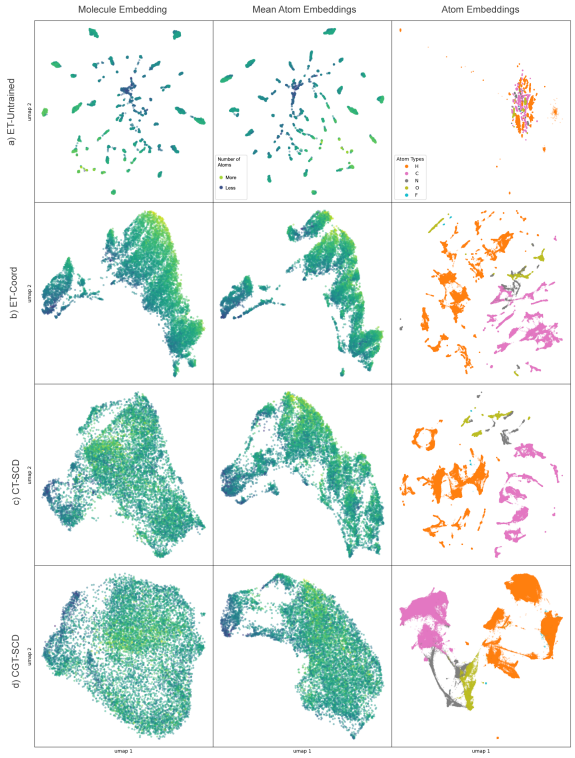

Figure 6: **Pre-training Effects on the Embedding Space.** UMAP visualizations of embeddings for 10k QM9 molecules, colored by atom count (N). **(a)** Untrained GNNs exhibit a strong bias towards "extensive" representations, clustering molecules primarily by size (N). **(b)** Standard node denoising smooths the latent manifold but embeddings remain highly extensive. **(c–d)** SCD pre-training yields a smoother, less extensive (more semantic) embedding space. We observe that downstream performance improves as representations become smoother and less extensive (Performance: $a < b < c < d$). 'Molecule embeddings' are created by passing sum pooled atom embeddings through a two-layer MLP 'embedding head'. In SCD 'Molecule embeddings' are used for self-embeddings. In our coord model, the molecule 'embedding head' remains untrained.

D.2 SUMMARY TABLES FOR THE MAIN-TEXT FIGURES

Table 11 lists the absolute errors plotted in Figure 3; Table 12 lists compute and all ten QM9 targets for the models in Figure 1.

Table 11: **Comparing the Impact of Different Pre-training Data**. Pre-training on any data source benefits every task, though downstream performance gains are greatest when pre-training on in-domain data. Datasets with high elemental diversity yield the highest average improvements. *CT-FE* models denote supervised force-energy pre-training; *CT-SCD* denotes self-supervised pre-training. Best results are **bold**; second best are underlined. Plotted in Figure 3.

	QM9 (Small Mol.)		Matbench (Crystals)		LBA (Proteins)		
Model (Pre-training Data)	**HOMO** (meV)	**% imp.**	**MP Gap** (eV)	**% imp.**	**RMSE** (id60)	**% imp.**	**Avg % imp.**
ET/CT (Baseline)	20.3	0.0%	0.186	0.0%	1.386	0.0%	0.0%
CT-SCD-PCQ	**12.7**	37.4%	0.174	6.5%	1.226	11.5%	18.5%
CT-SCD-GEOM10	13.2	35.0%	0.177	4.8%	1.211	12.6%	17.5%
CT-SCD-AMP20	14.6	28.1%	**0.122**	34.4%	1.219	12.0%	24.8%
CT-SCD-SAIR	14.6	28.1%	0.182	2.2%	1.196	13.7%	14.6%
CT-SCD-ALL	**12.7**	37.4%	0.132	29.0%	1.218	12.1%	26.2%
CT-FE-OMol25	13.6	33.0%	0.134	28.0%	1.187	14.4%	25.1%
CT-SCD-OMol25	12.8	36.9%	0.136	26.9%	**1.175**	15.2%	**26.4%**

Table 12: **Comparing Computational Costs and Performance**. SCD pre-training enables competitive performance at a lower cost than more advanced Clebsch-Gordan tensor product based architectures, and large-scale force-energy (FE) pre-trained models. **Bold** denotes best; underline denotes second best. Plotted in Figure 1.

Metric	**EquiformerV2**	**Baseline (ET)**	**CT-SCD**	**CGT-SCD**	**JMP-S**	**JMP-L**
Model Parameters	11.2M	7M	10M	17M	∼30M	235M
Pre-training Structures	0	0	3.4M (U)	3.4M (U)	120M (L)	120M (L)
Pre-training GPU-hours	0	0	**298**	1,308	5,700	34,400
Fine-tuning GPU-h/task	137	—	**46**	96	—	—
Total hours (10 tasks)	1,370	—	**758**	2,268	$> 5,700$	$> 34,400$
QM9 Task (MAE)						
HOMO (meV)	14.0	20.3	12.7	9.65	11.1	**8.8**
LUMO (meV)	13.0	17.5	11.5	9.05	10.8	**8.6**
Gap (meV)	29.0	36.1	24.5	19.7	23.1	**19.1**
ZPVE (meV)	1.47	1.84	1.18	1.22	1.0	**0.9**
α (a_0^3)	0.050	0.059	0.0377	0.0383	0.037	**0.032**
C_v (cal/mol K)	0.023	0.026	0.021	0.019	0.018	**0.017**
U_0 (meV)	6.17	6.15	3.58	3.96	3.3	**2.9**
U (meV)	6.49	6.38	3.50	4.11	3.3	**2.8**
H (meV)	6.22	6.16	3.52	4.00	3.3	**2.8**
G (meV)	7.57	7.62	5.29	5.29	4.5	**4.3**

(U) denotes unlabeled geometries; (L) denotes labeled force-energy structures. SCD GPU-hours were measured on NVIDIA A100 GPUs at fp32 without compilation; EquiformerV2 and JMP hours are as reported by their authors on different hardware.

D.3 Ligand Binding Affinity

D.3.1 Conditional Denoising of Coupled Components: Pockets and Ligands

When dataset samples ($\mathbf{x}$) contain interdependent pairs ($\mathbf{x} = A \cup B$), such as a protein pocket and a bound ligand, we can alter the self-conditioned denoising objective to instead denoise one component based on an embedding from the other (Equation 4). In doing so we can exploit known conditional relationships in the dataset during pre-training.

$$\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\|\phi_\theta(\tilde{\mathbf{x}}|\mathbf{c}) - \boldsymbol{\varepsilon}\|_2^2\right] = \mathbb{E}_{q_\sigma(\tilde{\mathbf{A}},\mathbf{B})}\left[\left\|\phi_\theta(\tilde{\mathbf{A}}|\phi_\theta(\mathbf{B})_{L0})_{L1} - \boldsymbol{\varepsilon}_A\right\|_2^2\right] \tag{4}$$

Table 13: **SCD on Proteins**: Ligand Binding Affinity (LBA) predictions over two sequence identity splits (30% and 60%). SCD models pre-trained on synthetic or diverse data (e.g., SAIR, OMol25) can outperform larger models pre-trained on PDB-based datasets. Best results within each category are **bolded**; second best are underlined.

Method	Model	Seq. overlap 30%			Seq. overlap 60%			Params
		RMSE ↓	Pear. ↑	Spear. ↑	RMSE ↓	Pear. ↑	Spear. ↑	
No Pre-training	HoloProt-Full Surface (Somnath et al., 2021)	1.464	0.509	0.500	1.365	0.749	0.742	1.4M
	ProtNet-All-Atom (Wang et al., 2023a)	1.463	0.551	0.551	1.343	0.765	0.761	—
	ATOM3D-3DCNN (Townshend et al., 2022)	1.416	0.550	0.553	1.621	0.608	0.615	—
	ATOM3D-ENN (Townshend et al., 2022)	1.568	0.389	0.408	1.620	0.623	0.633	—
	ATOM3D-GNN (Townshend et al., 2022)	1.601	0.545	0.533	1.408	0.743	0.743	—
	EPT-Scratch (Jiao et al., 2025)	**1.378**	**0.604**	**0.594**	**1.277**	**0.787**	**0.785**	30M
	CT	1.510	0.501	0.486	1.386	0.736	0.720	10M
Pre-trained on PCQ	Frad (Ni et al., 2024)	1.365	0.599	0.577	**1.213**	**0.804**	**0.801**	14M
	CT-SCD-PCQ	**1.332**	**0.617**	**0.600**	1.226	0.800	0.794	10M
Other Pre-trained	EGNN-PLM (Wu et al., 2023)	1.403	0.565	0.544	1.559	0.644	0.646	650M
	Uni-Mol (Zhou et al., 2023)	1.520	0.558	0.540	1.619	0.645	0.653	47.6M
	ProFSA (Gao et al., 2023)	1.377	0.628	0.620	1.377	0.764	0.762	47.6M
	EPT-Molecule (Jiao et al., 2025)	1.336	0.621	0.602	1.243	0.802	0.800	30M
	EPT-Protein (Jiao et al., 2025)	1.329	0.628	0.613	1.235	0.804	0.800	30M
	EPT-MultiDomain (Jiao et al., 2025)	1.322	0.644	0.630	**1.227**	**0.811**	**0.803**	30M
	ADiT-S (Shi et al., 2026)	1.337	0.626	0.618	1.413	0.740	0.740	12M
	ADiT-M (Shi et al., 2026)	1.353	0.622	0.630	1.335	0.764	0.752	35M
	ADiT-L (Shi et al., 2026)	**1.308**	**0.645**	**0.647**	1.246	0.767	0.765	253M
SCD Pre-trained	CT-SCD-AMP20	1.408	0.584	0.533	1.219	0.803	0.800	10M
	CT-SCD-GEOM10	1.392	0.606	0.554	1.211	0.806	0.801	10M
	CT-SCD-ALL	1.372	0.594	0.578	1.218	0.802	0.798	10M
	CT-SCD-SAIR-Lig	1.412	0.566	0.547	1.283	0.779	0.770	10M
	CT-SCD-SAIR	1.337	0.617	0.599	1.196	0.810	0.802	10M
	CT-SCD-SAIR-Pocket	**1.304**	**0.640**	**0.624**	1.200	0.809	0.806	10M
	CT-SCD-OMol25	1.389	0.586	0.571	**1.175**	**0.817**	**0.816**	10M
FE Pre-trained	CT-FE-OMol25	1.391	0.575	0.564	1.187	0.814	0.811	10M

In Table 13 we present results from three variations of pre-training on SAIR, illustrated in Figure 4. In CT-SCD-SAIR, we pre-train the model using canonical SCD, embedding and denoising the entire pocket-ligand complexes together. However, in pre-training CT-SCD-SAIR-Lig the ligand structure is embedded and used to conditionally denoise the pocket atoms; and in CT-SCD-SAIR-Pocket we do the reverse, pocket atoms are embedded and used to conditionally denoise the ligand atoms. In doing this we find that CT-SCD-SAIR-Pocket performs the best (specifically on low sequence overlap), CT-SCD-SAIR the second best, and CT-SCD-SAIR-Lig the worst. This follows biochemical intuition, as we would expect ligand conformation to be strongly dependent on pocket geometry, but not the other way around. We hold this as an example of how exploiting known conditional relationships can benefit learned representation during pre-training. We expect component-conditional denoising to transfer to other coupled systems with a known direction of dependence: adsorbates on surfaces, solutes in solvent shells, and dopants or defects in a host lattice, where in each case the environment is the natural conditioning component.

D.4 Force-Energy Pre-Training and Fine-Tuning for Property Prediction

In this work, we conduct all force-energy pre-training using the OMol25-4M dataset and a CT backbone. In Table 15, we compare models pre-trained via either direct ('forward') or gradient-based ('backward' aka 'conservative') force prediction. Unsurprisingly, the simple CT backbone does not yield state-of-the-art force-energy predictions; however, we find that SCD pre-training confers a modest benefit even when restricted to the same training geometries as the labeled task. Additionally, we find that a single SCD step (comprising a double forward pass) is roughly half the speed of a direct force training step (a single forward pass), but $1.4\times$ faster than a gradient-based (forward then 'backwards') force training update step. Table 16 presents benchmarking results for these models on two representative QM9 tasks: HOMO and U_0 (potential energy at 0K). We observe that pre-training via direct force prediction benefits downstream property prediction more than gradient-based force pre-training, but it does not outperform SCD pre-training. In main text tables, we report only the best model – pre-trained exclusively on direct force prediction – as 'CT-FE-OMol25'.

Table 14: **SCD on Periodic Materials**: Matbench band gap prediction performance for SCD and supervised force-energy (FE) pre-trained models. Best results within each category are **bolded**; second best are underlined. SCD models achieve competitive performance while utilizing significantly fewer parameters and unlabeled structures.

Category	Model	MP Gap (eV)		# Params	Pre-training
		Fold 0	Mean (0–4)		Data Size
No Pre-training	MODNet (De Breuck et al., 2021)	0.215	0.220	—	n/a
	coGN (Ruff et al., 2023)	**0.153**	**0.156**	—	n/a
	JMP-S (Shoghi et al., 2023)	0.235	—	∼30M	n/a
	JMP-L (Shoghi et al., 2023)	0.228	—	235M	n/a
	CT (Baseline)	0.186	—	10M	n/a
FE Pre-trained	JMP-L (Shoghi et al., 2023)	**0.089**	**0.091**	235M	120M
	JMP-S (Shoghi et al., 2023)	0.119	0.121	∼30M	120M
	HackNIP (Orb-v2 + MODNet) (Kim et al., 2025)	—	0.150	> 25.2M	> 32M*
	CT-FE-OMol25	0.134	—	10M	4M
Other SSL	Crystal-Twins (Magar et al., 2022)	—	0.264	—	428k
SCD Pre-trained	CT-SCD-AMP20	**0.122**	**0.123**	10M	675k
	CT-SCD-ALL	0.132	—	10M	11.3M
	CT-SCD-PCQ	0.174	—	10M	3.4M
	CT-SCD-GEOM10	0.177	—	10M	2.8M
	CT-SCD-SAIR	0.182	—	10M	4.4M
	CT-SCD-OMol25	0.136	—	10M	4M

*Orb-v2 data size estimated from the sum of force-energy fine-tuning datasets: MPTraj (1.58M) and Alexandria (30.5M) (Kim et al., 2025).

Table 15: **OMol25 4M Force-Energy Pre-training**. Comparison of supervised force-energy training baselines (FE) against fine-tuning from SCD pre-trained weights (SCD-FE). We compare direct force prediction (Forward) against gradient-based force calculation (Backward). CT-SCD-OMol25 is pre-trained by SCD using the same OMol25 4M geometries.

Model Name	Initialization	Force Method	Val Force MAE (meV/Å)	Val Energy MAE (meV)	num training steps
FE-Fwd	From Scratch	Forward	44.8	648	800k
FE-Fwd-long	From Scratch	Forward	42.6	606	1600k
FE-Bwd	From Scratch	Backward	50.0	618	800k
SCD-FE-Fwd	CT-SCD-OMol25	Forward	42.9 (+4.2%)	617 (+4.8%)	800k (+800k SCD)
SCD-FE-Bwd	CT-SCD-OMol25	Backward	41.4 (+17.2%)	665 (-7.6%)	800k (+800k SCD)

D.5 MD17: Small Molecule Forces

Consistent with prior work, we benchmark a smaller SCD model (6 layers, dim 128) on MD17 (Chmiela et al., 2017). The models presented in Table 18 were pre-trained on the PCQ dataset and subsequently fine-tuned on MD17. To match the fine-tuning methods described by Frad (Ni et al., 2024), we also train a model using a two pass 'force-denoise' scheme, in which the first pass is used to predict forces and energies and the second forward pass is used for an auxiliary denoising task. We observe that SCD pre-training with the CT backbone yielded results similar to those of Frad and SliDe, while achieving $4\times$ faster inference speeds. However, none of these methods achieve particularly strong performance on this benchmark. We hypothesize that this stems from the pre-training dataset containing exclusively ground-state geometries; consequently, the non-equilibrium structures in MD17 are likely out-of-distribution. Lacking non-equilibrium structures during pre-training, the models fail to capture features necessary for non-equilibrium prediction tasks. This aligns with our findings on other benchmarks, where similarity between pre-training and fine-tuning data distributions is a strong predictor of performance. We do not report results for the GET/CGT variant due to persistent segmentation faults when computing gradient-based forces.

Table 16: **Fine-tuning Force-Energy Pre-trained Models**. Force-energy pre-trained models performed best when the energy head was reset and regularization noise was turned off during fine-tuning. Forward (non-conservative) force pre-training conveyed the best average results on downstream property prediction and is reported as 'CT-FE-OMol25' elsewhere.

Model	**Fine-tuning Protocol**	**HOMO** (meV)	U_0 (meV)	avg % impv. over baseline
baseline		20.3	6.15	—
FE-Fwd	Standard	15.9	12.9	-44.0%
FE-Fwd	Reset Head, $\sigma_{\text{reg}} = 0.005$	15.0	10.2	-19.9%
FE-Fwd	Reset Head, $\sigma_{\text{reg}} = 0$	13.6	3.40	38.9%
FE-Fwd-long	Reset Head, $\sigma_{\text{reg}} = 0$	13.7	3.47	38.0%
FE-Bwd	Reset Head, $\sigma_{\text{reg}} = 0$	14.5	3.36	37.0%
CT-SCD-OMol25	Standard	12.8	3.54	39.7%

Table 17: SCD vs. other SSL Methods using different backbones and pre-training data. Models CT and CGT below are both pre-trained on the PCQ dataset. Best results are **bolded**; second best are underlined.

QM9 Task	**CT** (Ours)	**CGT** (Ours)	**EPT-10** (Jiao et al., 2025)	**UniCorn** (Feng et al., 2024)
Params	10M	17M	30M	—
Pre-train Data	3.4M	3.4M	5.9M	15M
HOMO (meV)	12.7	**9.65**	15.2	13.0
LUMO (meV)	11.5	**9.05**	13.6	11.9
Gap (meV)	24.5	**19.7**	29.0	24.9
ZPVE (meV)	1.18	1.22	**1.11**	1.4
U_0 (meV)	**3.58**	3.96	5.44	3.99
U (meV)	**3.50**	4.11	5.54	3.95
H (meV)	**3.52**	4.00	5.42	3.94
G (meV)	5.29	5.29	6.37	**5.09**
α (a_0^3)	0.0377	0.0383	0.045	**0.036**
C_v (cal/mol K)	0.021	**0.019**	0.020	**0.019**

The pre-training datasets for EPT-10 and UniCorn include the 3.4M PCQ structures used by SCD.

Table 18: SCD achieves comparable performance on MD17 force–energy prediction to prior methods while using a smaller and faster architecture. Best results are in **bold**, second best are underlined.

Method	**Baseline**	**Coord**	**SCD**	**SCD (force-denoise)**	**Frad (force-denoise)**	**SliDe**
Architecture	ET-small	ET-small	CT-small	CT-small	GET-small	GET-small
Pre-training dataset	none	PCQ	PCQ	PCQ	PCQ	PCQ
Forces MAE (kcal/mol/Å)						
Aspirin	0.253	0.211	0.210	0.200	0.209	**0.174**
Benzene	0.196	**0.169**	0.171	0.175	0.199	**0.169**
Ethanol	0.109	0.096	0.120	0.116	0.091	**0.088**
Malonaldehyde	0.169	0.139	0.186	0.173	**0.142**	0.154
Naphthalene	0.061	0.053	0.041	**0.040**	0.053	0.048
Salicylic acid	0.129	0.109	0.101	**0.099**	0.108	0.101
Toluene	0.067	0.058	0.053	**0.051**	0.054	0.054
Uracil	0.095	0.074	0.077	**0.072**	0.076	0.083
Energy MAE (kcal/mol)						
Aspirin	0.123	–	0.0742	**0.0705**	–	–
Benzene	0.058	–	0.0274	**0.0211**	–	–
Ethanol	0.052	–	0.0286	**0.0280**	–	–
Malonaldehyde	0.077	–	0.0432	**0.0417**	–	–
Naphthalene	0.085	–	0.0163	**0.0156**	–	–
Salicylic acid	0.093	–	**0.0286**	0.0290	–	–
Toluene	0.074	–	**0.0181**	0.0184	–	–
Uracil	0.090	–	**0.0190**	0.0193	–	–

CGT results on MD17 are omitted due to an unresolved segmentation fault during backpropagation.

E DATASETS

E.1 PRE-TRAINING DATASETS

PCQ: The PCQM4Mv2 dataset is a curated subsample of the PubChemQC project containing ground state geometries for 3.378M organic small molecules obtained using B3LYP/6-31G* DFT (Nakata & Shimazaki, 2017). This dataset is commonly used for demonstrating self-supervised learning methods, either on its own or as part of a composite dataset. Throughout this work we use 'PCQ' to indicate models trained only on this dataset.

GEOM10: The original GEOM dataset contains 36M conformers from ∼450k unique molecules generated using GFN2-xTB, a semi-empirical extended tight-binding method (Axelrod & Gómez-Bombarelli, 2022). GEOM is split into three categories: 1) GEOM-drugs, containing conformers of drug like small molecules, 2) GEOM-QM9, conformers of molecules from the QM9 dataset, and 3) GEOM-MoleculeNet, containing conformers from a subsample of molecules in the MoleculeNet benchmark. To avoid overlap with benchmarking datasets we only utilize structures from GEOM-drugs, containing 304k unique molecules, and curate two subsamples: GEOM10, containing up to 10 of the lowest energy conformers from each molecule (2.7M geometries total), and GEOM1 containing only the lowest energy (ground-state) conformer from each molecule (304k geometries).

SAIR: The Structurally Augmented IC50 Repository (SAIR) dataset contains a large number of bound protein-ligand structures generated using the Boltz-1x model (Wohlwend et al., 2025). Five conformers are provided for each structure. Although the authors of SAIR report 1,048,857 unique binding pairs totaling 5.2M conformers, the copy we obtained from Hugging Face (`https://huggingface.co/datasets/SandboxAQ/SAIR`) contains only 888,104 unique binding pairs and 4.4M conformers.

Alex-MP-20: The Alex-MP-20 dataset comprises 675,204 periodic materials structures with up to 20 atoms (Zeni et al., 2023) derived from both the Materials Project (Jain et al., 2013), and the Alexandria materials database (Schmidt et al., 2023; 2022). Throughout this work, we refer to Alex-MP-20 as 'AMP20'. When SCD pre-training on AMP20, we apply a unit cell repeat augmentation to simulate larger crystals. Each sample is randomly tiled in one direction with a probability ($p_{\text{cell repeat}}$) of 50%, for up to two repeats. This dataset is also available on Hugging Face (`https://huggingface.co/datasets/OMatG/Alex-MP-20`).

AllAtoms (ALL): We also pre-train on a combined dataset that includes the entirety of PCQ, GEOM10, SAIR, and AMP20, totaling 5.2M unique structures with 11.3M conformations.

OMol25 4M split: The Open Molecules 2025 (OMol25) dataset is a large-scale repository of over 100 million quantum chemistry calculations covering 83 million unique molecular systems (Levine et al., 2025). While the full dataset spans the first 83 elements (H–Bi) and systems up to 350 atoms, we utilize the designated '4M split', a randomly sampled subset of approximately 4 million geometries. All labels (forces and energies) were computed at the ωB97M-V/def2-TZVPD level of theory. The 4M subset preserves the broad chemical diversity of the full dataset, containing equilibrium and non-equilibrium structures from four primary domains: small organic molecules, biomolecules (proteins, DNA/RNA from PDB/BioLiP2), metal complexes (transition metals, ligands), and electrolytes. This dataset is available on Hugging Face (`https://huggingface.co/facebook/OMol25`).

E.2 BENCHMARKING DATASETS AND FINE-TUNING HYPERPARAMETERS

QM9: The QM9 dataset is a commonly used benchmark for small molecule property prediction on 3D geometries (Ramakrishnan et al., 2014). QM9 comprises a set of small organic molecules containing up to nine heavy atoms (C, N, O, F), or up to 29 atoms total when including hydrogen. All 3D Geometries from QM9 are computed using the B3LYP/6-31G(2df, p) DFT functional and are labeled with a set of 15 quantum mechanical properties including HOMO-LUMO gap, atomization energies, specific heat, polarizability and more. The original dataset contains 134k molecules; however, it is now standard practice to use a cleaned version containing 130,831. We obtain our copy from the PyTorch Geometric library. QM9 is typically split randomly: 110,000 for training, 10,000 for validation, and the remaining 10,831 for the test set. We use this convention in our work; however, we do not use the validation set for early stopping. We follow the practices described in TorchMD-Net

(Thölke & De Fabritiis, 2022) and subtract provided reference energies from atomization energy targets. All other targets are normalized to a standard Gaussian for prediction.

LBA: The Ligand Binding Affinity (LBA) dataset comprises 4,463 co-crystallized 3D structures of protein-ligand complexes derived from PDBbind. Binding affinity targets are experimentally derived and provided as negative log of either inhibition or dissociation constant ($-\log(K)$). Two splits are provided based on sequence overlap between proteins in the training set and test set; in 'id60' no protein in the training set has more than a 60% sequence overlap with any protein in the test set. Similarly, in 'id30' train-test sequence overlap is limited to 30%. Following ATOM3D preprocessing (Townshend et al., 2022), predictions are made using extracted pocket-ligand complexes.

Matbench: The matbench properties dataset (Dunn et al., 2020) provides 13 different tasks (regression and classification) for periodic materials structures to benchmark property prediction models. Across these tasks, training set size ranges from 312–132k samples, each of which is split into 5 folds with predefined test splits. In this work we explore only one regression task, 'mp_gap' or bandgap energy prediction, as a proof of concept. We choose bandgap energy because of its economic relevance. The 'mp_gap' task provides 106k periodic crystal structures in total, resulting in training set sizes of 85k and test set sizes of 21k. We report results for split 0 with all models tested. For our best model we also report the average of all splits.

MD17: The MD17 (Chmiela et al., 2017) dataset comprises a collection of ab initio molecular dynamics (AIMD) trajectories for eight small organic molecules with DFT forces and energies, commonly used to evaluate MLIPs. Following past convention we use a random split of 950 samples for training, 50 for validation (however we do not employ early stopping), and the remainder as the test set. The total number of geometries provided for each trajectory varies from 150k-1M. All force predictions provided are derived from energy gradients (a backwards pass).

F Hyperparameters

F.1 Constants Across All Training Runs

All training runs discussed in this work use the AdamW optimizer with β_2=0.999, and a learning rate scheduler composed of a linear warm up, followed by cosine decay. Unlike many prior works, we do not use validation-based early stopping. All graph creation uses a cutoff distance of 5.0 Å. Tables 19 and 20 list the SCD pre-training hyperparameters; Tables 21, 22, 23 and 24 list the fine-tuning hyperparameters for QM9, Matbench, LBA and MD17.

Table 19: Base SCD pre-training hyperparameters: applied in all SCD training runs

Hyperparameter	**Value**
Warmup steps	10,000
Learning rate	0.005
β_1	0.9
Weight decay	0.05
DropPath	0.1
Corrupt noise σ	0.04
Regularization noise σ	0.005

Table 20: SCD Pre-training: dataset and model specific hyperparameters.

Backbone	CT	CGT	CT	CT	CT	CT	CT
Dataset	PCQ	PCQ	GEOM10	AMP20	SAIR	ALL	OMol25
Total steps	800k	800k	800k	800k	800k	1.2M	800k
Batch size / GPU	128	32	48	142	12	54	48
# GPUs (A100)	4	16	16	4	8	16	16
Effective batch size	512	512	768	568	96	864	768
Materials augmentations							
Cell repeat iters	–	–	–	2	–	2	–
$p_{\text{cell repeat}}$	–	–	–	0.5	–	0.5	–

Table 21: QM9 fine-tuning hyperparameters for different target groups.

Hyperparameter	**HOMO, LUMO, Gap**	U_0, U, H, G	α, **ZPVE,** C_v
Batch size	128	32	128
Learning rate	5e-4	3e-4	3e-4
Total steps	300k	500k	400k
Warmup steps	10k	10k	10k
β_1	0.995	0.995	0.95
Weight decay	0.01	0.01	0.01
DropPath (init / final)	0.15 / 0.1	0.15 / 0.0	0.15 / 0.01
Reg. noise σ	0.005	0.0	0.005
Denoise loss weight	0.1	0.0	0.2
Subtract ref. energy	No	Yes	No
Loss EMA	Off	0.05	Off
Head aggregation	Sum	Sum	Sum

Table 22: Matbench fine-tuning hyperparameters for the `mp_gap` task.

Hyperparameter	**Value**
Batch size	8
Learning rate (from scratch)	2e-4
Learning rate (fine-tuning)	8e-5
Total steps	300k
Warmup steps	10k
β_1	0.995
Weight decay	0.01
DropPath (init / final)	0.15 / 0.00
Reg. noise σ	5e-4
Denoise loss weight	0.01
Head aggregation	Sum

Table 23: Fine-tuning hyperparameters for the Ligand Binding Affinity (LBA) benchmark

Hyperparameter	**Value**
Batch size	8
Learning rate (from scratch)	2e-4
Learning rate (fine-tuning)	1e-4
Total steps	30k
Warmup steps	500
β_1	0.995
Weight decay	0.05
DropPath	0.15
Reg. noise σ	0.005
Denoise loss weight	0.1
Head aggregation	Mean

Table 24: Training hyperparameters for the MD17 benchmark.

Hyperparameter	**Value**
Backbone	CT-small
Training set size	950
Batch size	8
Energy loss EMA	0.05
Energy loss weight	0.8
Force loss weight	0.2
Learning rate	5e-4
Warmup steps	5k
Total steps	150k
β_1	0.995
Weight decay	0.01
DropPath (init / final)	0.15 / 0.00
Reg. noise σ	0.005
Denoise loss weight	0.1
Head aggregation	Sum