GeoRecon: Graph-Level Representation Learning for 3D Molecules via Reconstruction-Based Pretraining

Shaoheng Yan [1] **Zian Li** [1 2] **Muhan Zhang** [1]

Abstract

Pretraining—finetuning has driven major advances in natural language processing and vision through objectives such as masked language modeling and next-token prediction. In molecular representation learning, however, pretraining tasks remain largely restricted to *node-level* denoising, which captures local atomic environments but provides no explicit training signal for the global molecular structure needed by graph-level property prediction tasks such as energy estimation and molecular regression. To address this gap, we introduce GeoRecon, a *graph-level* pretraining framework that shifts the reconstruction target from individual atoms to the molecule as an integrated whole. During pretraining, GeoRecon learns a graph representation that conditions geometry reconstruction and induces smoother, more transferable latent spaces. This encourages coherent global structural features beyond isolated atomic details while remaining fully self-supervised and 3D-only. Across QM9, MD17, MD22, and appendix benchmarks, GeoRecon consistently improves over its direct coordinate-denoising baseline and remains competitive with broader molecular pretraining baselines, supporting graph-level reconstruction as a simple and effective complement to node-level denoising.

1. Introduction

With the pretraining–finetuning paradigm extending from computer vision and natural language processing (Bao et al., 2021; Lu et al., 2019) to an increasingly broad range of domains, designing effective pretraining tasks has become a key component of modern training pipelines. In molecular representation learning, denoising- and masking-based pretraining has been widely validated (Zaidi et al., 2023; Zhou et al., 2023; Feng et al., 2023). In these tasks, molecular structures are perturbed or masked at the *atomic (**node**) level*, and the model is trained to recover the original local information.

However, many downstream molecular tasks require strong *whole-molecule (**graph**) level* understanding (Li et al., 2020). Energy prediction, molecular regression, and force-field learning all depend on local geometry, but they also require coherent global context. Conventional node-level pretraining tasks do not explicitly align the representation with such graph-level objectives, which limits the gains that pretraining can provide for molecule-level prediction.

We conducted a preliminary experiment by perturbing the atomic coordinates of a molecule with small displacements and using a spectral norm method described in Appendix D to estimate the Lipschitz constant of the encoder in the public Coord (Zaidi et al., 2023) pretraining checkpoint, a standard node-denoising-based baseline. This constant measures the sensitivity of the encoder's output representation to coordinate changes (larger values indicate higher sensitivity).

Table 1. Local non-rigid Lipschitz constant $L(x)$ for GeoRecon and the Coord baseline under two parameter settings: finite-difference perturbation magnitude (`fd_eps`, left) and power-method iteration count (`step`, right).

Model	`fd_eps`	2×10^{-4}	1×10^{-4}	5×10^{-5}
GeoRecon	*median*	3.069×10^{1}	3.071×10^{1}	3.073×10^{1}
	p95	5.161×10^{1}	5.090×10^{1}	5.092×10^{1}
Coord	*median*	2.517×10^{4}	2.539×10^{4}	2.539×10^{4}
	p95	4.490×10^{4}	4.638×10^{4}	4.469×10^{4}

Model	`step`	5	15	25
GeoRecon	*median*	2.969×10^{1}	3.071×10^{1}	3.075×10^{1}
	p95	5.087×10^{1}	5.090×10^{1}	5.090×10^{1}
Coord	*median*	2.537×10^{4}	2.539×10^{4}	2.539×10^{4}
	p95	4.629×10^{4}	4.638×10^{4}	4.451×10^{4}

[1]Institute for Artificial Intelligence, Peking University; State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China [2]School of Intelligence Science and Technology, Peking University, Beijing, China. Correspondence to: Muhan Zhang <muhan@pku.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

As shown in the second row of Table 1, Coord exhibits values exceeding 10^4, indicating extreme sensitivity to minute perturbations. This instability underscores the limitations of node-level pretraining, as smooth representation manifolds are well known to facilitate downstream performance (Lee et al., 2025; Guo et al., 2024; Zhang et al., 2025; Krishnan et al., 2020). We provide a brief proof in Appendix E.

To address such limitations, prior work has introduced explicit supervision via graph-level attributes (Hu et al., 2020) or frequent functional motifs (Rong et al., 2020), or adopted contrastive learning (Liu et al., 2022). However, these methods are often over-specialized by hand-crafted supervision, limiting their general applicability, or are primarily designed for 2D or mixed 2D&3D settings, lacking the inherent support for fully 3D molecular pretraining offered by denoising-based frameworks (Zaidi et al., 2023).

To bridge this gap, we propose **Geo**metric **Recon**struction (GeoRecon), a graph-level 3D self-supervised pretraining task. Given a 3D molecular structure, the model is trained to *produce an informative, orientation-invariant graph-level representation* that guides reconstruction from strongly noised coordinates. This objective encourages the encoder to preserve molecule-level structural information that is useful for recovering perturbed geometries.

GeoRecon possesses several key properties that make it distinct from prior graph-level pretraining approaches and broadly applicable across molecular modeling scenarios:

- **3D-only input**: Relies *solely* on atomic coordinates, without requiring 2D bond graphs or structural annotations, avoiding dependence on ambiguous 2D labels;
- **Fully self-supervised**: Learns directly from raw coordinates without any external labels;
- **Label-free and broadly applicable**: Does not require additional annotations (e.g., spectral data), enabling application to a wide range of molecular datasets;
- **Architecture-agnostic pretraining objective**: Can be integrated into different SE(3)-equivariant backbones without altering downstream finetuning protocols.

To demonstrate the effectiveness of GeoRecon, we finetune the pretrained model on graph-level molecular property prediction benchmarks including QM9 (Ramakrishnan et al., 2014), MD17 (Chmiela et al., 2017), and the challenging MD22 (Chmiela et al., 2023) dataset. GeoRecon consistently improves over the direct coordinate-denoising baseline on MD17 and MD22 and improves most QM9 targets.

We further assess robustness and generality through appendix evaluations on training set size, encoder depth, pooling strategy, decoder depth, reconstruction noise scale, larger recent datasets, and out-of-distribution 3BPA. Taken together, these results indicate that explicit graph-level reconstruction can improve accuracy and robustness, while also clarifying where gains are modest and where additional matched comparisons remain useful.

2. Related Work

2D Molecular Graph Models. A large body of molecular representation learning treats molecules as 2D graphs, where atoms are nodes and chemical bonds are edges. The message passing neural network (MPNN) framework (Gilmer et al., 2017) provides a general formulation for learning atom–bond interactions through iterative neighborhood aggregation, followed by permutation-invariant graph pooling for molecular property prediction. Subsequent molecular GNNs improve this paradigm with chemistry-aware message passing. For example, D-MPNN propagates messages over directed bonds to reduce redundant backtracking and achieves strong performance across molecular property benchmarks (Yang et al., 2019).

Beyond supervised learning, self-supervised 2D molecular graph pretraining has been widely explored to reduce label dependence. Early strategies combine node- and graph-level objectives such as attribute masking and context prediction (Hu et al., 2020). GROVER (Rong et al., 2020) scales graph Transformer pretraining to large molecular corpora with node-, edge-, and graph-level tasks. GraphMAE (Hou et al., 2022) adopts masked graph autoencoding, while Mole-BERT (Xia et al., 2023) rethinks atom tokenization and masked modeling for molecular graphs. These approaches establish strong topology-based baselines, but 2D graphs cannot directly encode conformational geometry, distances, bond angles, torsions, or equivariant physical constraints crucial for quantum-chemical properties. This limitation motivates either augmenting 2D encoders with generated conformers or developing 3D-native molecular representation learning.

3D Geometric Graph Neural Networks. To incorporate spatial information, geometric GNNs operate on molecular conformations and model interactions in 3D space. SchNet (Schütt et al., 2018) uses continuous-filter convolutions over Cartesian coordinates to predict energies and forces. DimeNet (Gasteiger et al., 2020b) and DimeNet++ (Gasteiger et al., 2020a) introduce directional message passing to capture angular information, while GemNet (Gasteiger et al., 2021) further improves expressiveness through higher-order directional interactions. Another line enforces Euclidean symmetry through equivariant architectures such as EGNN (Satorras et al., 2021) and SE(3)-Transformer (Fuchs et al., 2020). These models provide expressive 3D backbones, but architecture design alone cannot fully address scarce labeled quantum-chemical data,

making self-supervised 3D pretraining important.

Modern Denoising Approaches. Coordinate denoising has become a principled self-supervised objective for 3D molecular pretraining. Coord pretraining (Zaidi et al., 2023) shows that recovering equilibrium conformations from Gaussian perturbations is closely connected to learning molecular force fields, providing useful supervision from unlabeled structures. Recent denoising methods refine the perturbation distribution to better approximate physically meaningful molecular motions. Frad (Feng et al., 2023) combines dihedral-angle and Cartesian perturbations, but denoises only the coordinate fraction to preserve force-learning equivalence while improving low-energy conformer coverage. SliDe (Ni et al., 2024) derives perturbations from classical intramolecular potential terms by perturbing bond lengths, angles, and torsions, and uses random slicing to avoid expensive Jacobian computation. DenoiseVAE (Liu et al., 2025b) learns molecule- and atom-adaptive noise distributions with a VAE-style generator, addressing shared hand-designed noise schedules. AniDS (Liu et al., 2025a) further moves from isotropic or homoscedastic perturbations to atom-specific anisotropic Gaussian noise with SO(3)-equivariant covariance modeling. These methods strengthen local geometric and force-field supervision, but their objectives mainly operate at the atom or coordinate level and do not explicitly require graph representations to encode molecule-wide coherence.

Graph-Level Molecular Pretraining and Positioning of GeoRecon. Graph-level objectives capture molecule-wide semantics beyond local atom-level patterns. In 2D molecular pretraining, motif prediction, motif generation, and graph-level contrastive learning encourage global consistency (Rong et al., 2020; Zhang et al., 2021; You et al., 2020; Qiu et al., 2020). Cross-modal methods further align 2D topology with 3D geometry: GraphMVP (Liu et al., 2022) aligns 2D graphs and 3D conformers through contrastive and generative objectives, MoleculeSDE (Liu et al., 2023a) uses group-symmetric stochastic differential equations to connect 2D and 3D modalities, Uni-Mol2 (Ji et al., 2024) jointly encodes atomic, graph, and spatial features at scale, and MolGT (Chen et al., 2025) combines node- and graph-level pretext tasks across 2D topology and 3D geometry. While effective, these approaches often require additional modalities, multi-view alignment, augmentation pipelines, or complex training designs.

3. Preliminaries

Denoising has emerged as a principled and empirically effective pretraining objective in 3D molecular representation learning (Zaidi et al., 2023; Feng et al., 2023; Zhou et al., 2023). Formally, given a slightly perturbed molecular conformation $\widetilde{\boldsymbol{r}} = \boldsymbol{r} + \boldsymbol{\epsilon} \in \mathbb{R}^{n\times 3}$, where n is the number of atoms, $\boldsymbol{r} \in \mathbb{R}^{n\times 3}$ is an equilibrium structure, and $\boldsymbol{\epsilon}$ is Gaussian noise, the model is asked to predict the noise from the perturbed conformation, $\mathrm{GNN}(\widetilde{\boldsymbol{r}}) \approx \boldsymbol{\epsilon}$. Zaidi et al. (2023) formally established the theoretical equivalence between coordinate denoising and force field prediction by leveraging the connection to score matching.

Theorem 3.1 (Equivalence between denoising and force field prediction (Zaidi et al., 2023))**.**

$$\mathbb{E}_{q_\sigma(\widetilde{\boldsymbol{r}},\boldsymbol{r})}\left[\left\|\mathrm{GNN}_\theta(\widetilde{\boldsymbol{r}}) - \nabla_{\widetilde{\boldsymbol{r}}}\log q_\sigma(\widetilde{\boldsymbol{r}} \mid \boldsymbol{r})\right\|^2\right] = \mathbb{E}_{q_\sigma(\widetilde{\boldsymbol{r}},\boldsymbol{r})}\left[\left\|\mathrm{GNN}_\theta(\widetilde{\boldsymbol{r}}) - \frac{\boldsymbol{r}-\widetilde{\boldsymbol{r}}}{\sigma^2}\right\|^2\right].$$

$\boldsymbol{r}$ denotes the equilibrium molecular structure, $\widetilde{\boldsymbol{r}}$ is its perturbed version obtained by Gaussian corruption, and θ represents the parameters of the denoising network. $q_\sigma(\widetilde{\boldsymbol{r}} \mid \boldsymbol{r})$ is defined as a Gaussian centered at $\boldsymbol{r}$, serving as a tractable approximation to the Boltzmann distribution $p_{physics}(\widetilde{\boldsymbol{r}}) \propto \exp(-E(\widetilde{\boldsymbol{r}}))$. Its score function corresponds to the standardized noise vector $(\boldsymbol{r}-\widetilde{\boldsymbol{r}})/\sigma^2$. More broadly, $p_{physics}$ can be approximated by a Gaussian mixture $q_\sigma(\widetilde{\boldsymbol{r}}) \sim \frac{1}{n}\sum_i \mathcal{N}(\widetilde{\boldsymbol{r}}; \boldsymbol{r}_i, \sigma^2 I)$, whose score defines a force field over noisy conformations.

This theoretical result provides a rigorous foundation for coordinate denoising as a physically grounded pretraining objective, and is empirically validated by the Coord model.

Building upon coordinate denoising, Feng et al. (2023) introduced additional dihedral angle noise $\delta\psi \sim \mathcal{N}(0, \sigma^2 I_m)$ along rotatable bonds. While Feng et al. (2023) advance coordinate denoising by incorporating molecular flexibility through dihedral perturbations, their framework still focuses on node-level supervision. The model predicts local coordinate noise for each atom without an explicit mechanism requiring the representation to encode global structural context.

4. Our Method

Building on the self-supervised, physically grounded coordinate-denoising objective (Zaidi et al., 2023; Feng et al., 2023), we exploit its dual benefit as both a force-field surrogate and a fine-grained geometric bias. Yet, when used in isolation, this purely node-level signal cannot enforce the global structural coherence required by graph-level property prediction. This gap motivates a complementary mechanism that explicitly encourages the encoder to capture molecule-wide dependencies rather than only local perturbations.

Design principles.

- **P1 (reconstruction as denoising).** We formulate reconstruction as predicting synthetic coordinate noise,

avoiding orientation ambiguity that arises when decoding absolute coordinates from an invariant graph embedding.

- **P2 (weak decoder).** We use a lightweight reconstruction decoder that is too weak to solve the task via local smoothing alone, thereby forcing the model to leverage global information.
- **P3 (graph-level conditioning).** We condition reconstruction by broadcasting a *pooled* graph embedding $\boldsymbol{g}$ (rather than clean node embeddings), so $\boldsymbol{g}$ is forced to encode molecule-level semantics rather than directly memorizing clean node-level details, while still providing sufficient signal for atom-wise noise prediction.

We therefore introduce GeoRecon, a reconstruction-based pretraining framework designed to capture global molecular structure from noisy inputs. A naive approach would directly decode coordinates from pooled graph-level embeddings, but this loses spatial granularity and creates orientation ambiguity when invariant graph embeddings are asked to predict equivariant coordinates. We instead formulate reconstruction as denoising over entire molecular geometries: we inject scaled Gaussian noise into the 3D coordinates and train a lightweight decoder, *conditioned on graph-level representations*, to predict the scaled noise.

More precisely, GeoRecon extends standard coordinate denoising by increasing the task difficulty and conditioning the decoder on graph-level representations. Unlike approaches that adapt noise to specific physical semantics (Feng et al., 2023; Liu et al., 2025b), we deliberately use isotropic coordinate noise and reduce decoder capacity. These design choices reduce reliance on local shortcuts and increase the role of the pooled global embedding, encouraging the encoder to capture long-range dependencies and coherent molecule-level geometry.

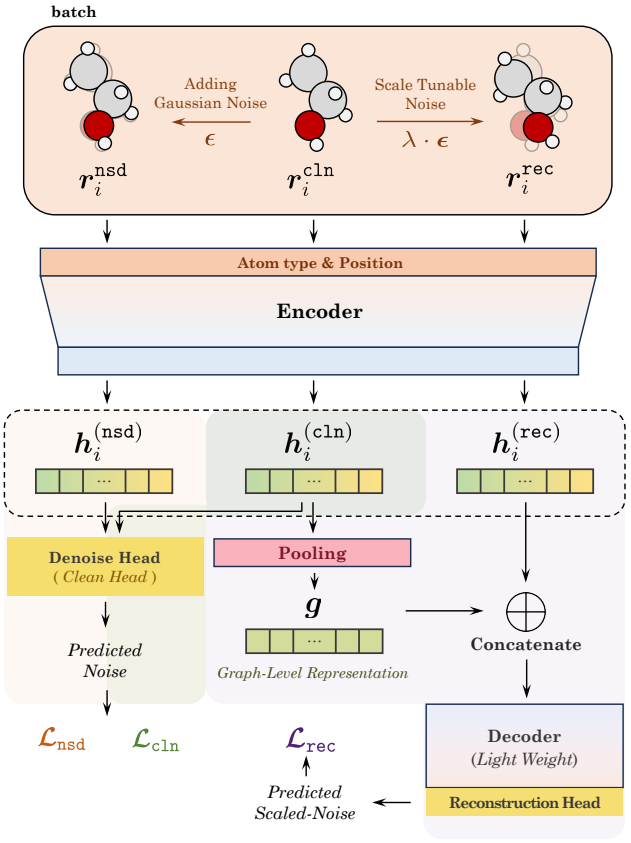

Figure 1. Overview of the GeoRecon framework. Given a molecular structure with atom types and 3D coordinates, the model encodes it using SE(3)-equivariant attention. Besides the standard node-level denoising objective, GeoRecon feeds a pooled graph-level representation concatenated with node embeddings derived from noisy coordinates into a lightweight decoder to reconstruct the additional noise. The pretrained encoder is then finetuned for downstream molecular property prediction tasks.

4.1. Notations

To support multi-task supervision, we construct three coordinate variants per molecule. (i) POS_{CLN}: the original coordinates $\boldsymbol{r}_i^{\mathtt{cln}} \in \mathbb{R}^3$, which provide the ground truth for denoising and are also used to compute the clean graph embedding $\boldsymbol{g}$; (ii) $\text{POS}_{\text{NOISED}}$: the perturbed coordinates $\boldsymbol{r}_i^{\mathtt{nsd}} \in \mathbb{R}^3$ used as input for the standard node-level denoising task; (iii) POS_{REC}: coordinates $\boldsymbol{r}_i^{\mathtt{rec}} \in \mathbb{R}^3$ corrupted with an alternative noise scale λ, used as the input to the reconstruction-as-denoising task.

The injected noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$ is shared across these variants. Specifically, we define:

$$\boldsymbol{r}_i^{\mathtt{nsd}} = \boldsymbol{r}_i^{\mathtt{cln}} + \boldsymbol{\epsilon}, \quad \boldsymbol{r}_i^{\mathtt{rec}} = \boldsymbol{r}_i^{\mathtt{cln}} + \lambda \cdot \boldsymbol{\epsilon}$$

where λ modulates reconstruction difficulty (larger λ means stronger perturbations), and is set to 1.35 by ablation. In particular, *scaled noise prediction* means predicting the perturbation generating POS_{REC}, i.e., the supervision target is $\lambda \cdot \boldsymbol{\epsilon}$. Each coordinate is paired with its corresponding atomic number z_i to construct the initial atomic embeddings.

Node-level embeddings are denoted as $\{\mathbf{h}_i^{(*)}\}_{i=1}^N$, where $* \in \{\mathtt{nsd}, \mathtt{rec}, \mathtt{cln}\}$ indicates the input variant corresponding to each pretraining task. The graph-level representation is denoted by $\mathbf{g} \in \mathbb{R}^d$, where d is the embedding dimension.

The model is trained to predict noise vectors $\hat{\boldsymbol{\epsilon}}_i^{(*)} \in \mathbb{R}^3$ for each atom and task, where the supervision signal is the synthetic Gaussian noise applied during pretraining. We denote the loss terms by $\mathcal{L}_*$, where the subscript indicates the task name, and use λ_* for the corresponding weighting coefficients.

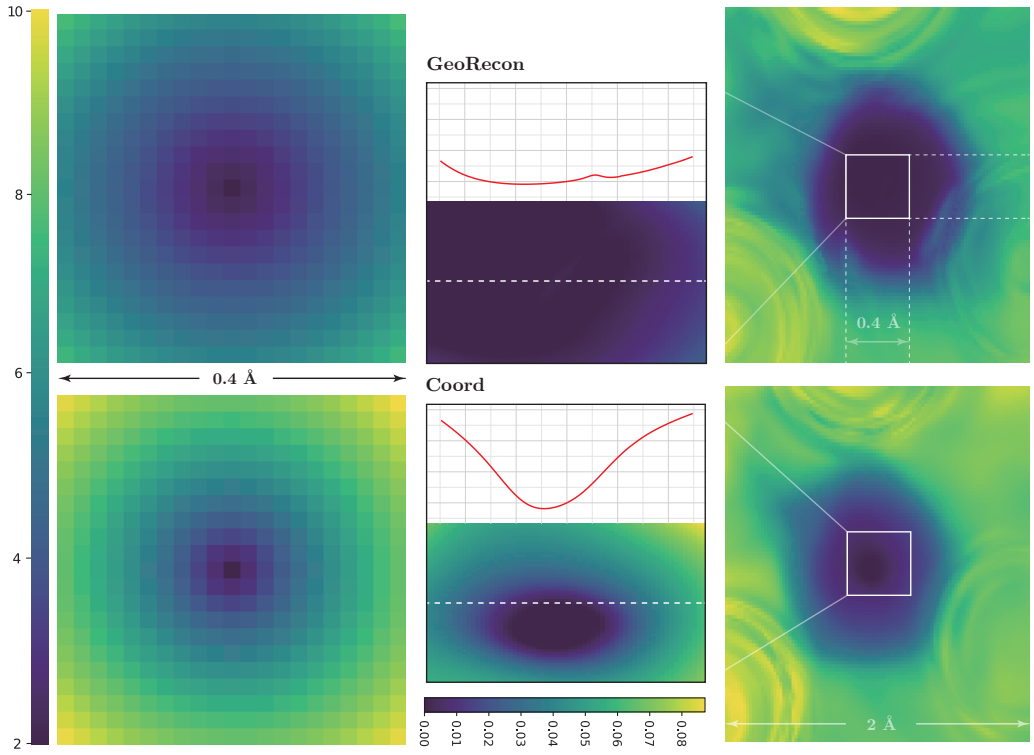

Figure 2. Representation stability of GeoRecon (upper row) vs. Coord (lower row) near equilibrium conformations ($\|\delta \boldsymbol{x}\| \leq 1$ Å). **Left**: averages over molecules. **Middle & Right**: 2D perturbation heatmaps for a randomly selected PCQM4Mv2 sample, where the horizontal and vertical axes correspond to magnitudes of two coordinate perturbations applied to a random atom, and the color encodes the norm change of the representation ($\|\delta \boldsymbol{h}\|$). Larger blue regions indicate higher representation stability near the equilibrium conformation. The red curve in the middle column highlights $\delta y = 0$ ($\|\delta \boldsymbol{h}\|$ along the x-axis). **Scale bars**: shared on the left (left) and at the bottom (middle).

4.2. Model Architecture

Shared Equivariant Encoder At the core of GeoRecon is an SE(3)-equivariant transformer encoder adapted from TorchMD-Net (Thölke & De Fabritiis, 2022). The encoder operates on molecular graphs defined by atomic numbers $\{z_i\}$ and spatial coordinates $\{\boldsymbol{r}_i^*\}$.

Denoising Head The denoising head reverses the synthetic perturbation applied during pretraining. We follow the node-level denoising paradigm for consistency with prior work and ease of integration, while adding objectives that explicitly require global graph structure.

According to Theorem 3.1, the equivalence between coordinate denoising and force field learning under the Boltzmann distribution assumption can be readily established. A brief derivation is included in the Appendix F for completeness and reference.

To fully leverage the clean molecular conformations already processed during pretraining, we integrate them into the denoising pipeline as well. Specifically, clean molecules, with zero perturbation, are passed through the same encoder and denoising head. Their supervision target is a zero noise vector, naturally enforcing representational consistency between clean and noised structures. This unifies the treatment of clean and noised molecules under a single objective, regularizing the encoder without requiring additional components or losses. To evaluate the impact of the auxiliary clean alignment task, we conduct additional ablation studies in Appendix G.

Reconstruction Head with Graph-Level Conditioning As mentioned previously, instead of directly decoding absolute coordinates, GeoRecon formulates reconstruction as a denoising task with stronger perturbations: the model takes POS_{REC} as input and predicts the corresponding (scaled) noise. The decoder is deliberately lightweight and is conditioned on a graph-level embedding.

To construct the conditioning signal, the clean input undergoes a complete forward pass through the encoder, producing node-level representations $\{\boldsymbol{h}_i^{(\texttt{cln})}\}$. Using the clean conformation keeps the global conditioning signal stable rather than dominated by the strong reconstruction noise. These representations are aggregated with a pooling function to yield a graph-level vector. The resulting vector $\boldsymbol{g}$ encapsulates holistic structural semantics and is broadcasted

to all nodes in the reconstruction input. We adopt mean pooling for its simplicity and stability under coordinate perturbations, while noting that more sophisticated pooling strategies are also compatible with our framework; see Appendix O for further discussion. Each node concatenates this global vector with its local representation $\boldsymbol{h}_i^{(\mathtt{rec})}$, forming the input to a lightweight decoder:

$$\boldsymbol{z}_i = \textsc{Concat}(\boldsymbol{g}, \boldsymbol{h}_i^{(\mathtt{rec})}), \quad \hat{\boldsymbol{\epsilon}}_i^{(\mathtt{rec})} = \mathrm{MLP}_{\mathtt{rec}}(\boldsymbol{z}_i).$$

Consequently, the learned latent space remains *locally descriptive*, capturing fine-grained atomic perturbations, while also being *globally coherent* with the overall molecular geometry. This property is useful for downstream molecular property prediction, especially for tasks that depend strongly on spatial organization, including total energy, dipole moment, and orbital energy estimation.

4.3. Multi-Task Training Objective

To jointly optimize local and global aspects of molecular geometry, we combine the three pretraining objectives into a unified loss:

$$\mathcal{L}_{\mathtt{total}} = \lambda_{\mathtt{nsd}}\mathcal{L}_{\mathtt{nsd}} + \lambda_{\mathtt{rec}}\mathcal{L}_{\mathtt{rec}} + \lambda_{\mathtt{cln}}\mathcal{L}_{\mathtt{cln}}.$$

Each loss term provides complementary supervision, encouraging the encoder to learn representations that are both locally descriptive and globally coherent. The loss weights are empirically chosen to balance the magnitudes of individual terms during training.

Taken together, these design choices aim to (i) reduce representation sensitivity to small coordinate perturbations by anchoring learning on a stable graph-level signal, and (ii) improve graph-level downstream generalization while keeping the pretraining recipe close to coordinate denoising. We test this causal chain in the experiments by first measuring representation stability, then evaluating downstream performance on QM9/MD17/MD22, and finally probing representation quality under a frozen-encoder linear probing setting.

5. Experiments

We evaluate GeoRecon under a controlled pretraining–finetuning protocol. Unless otherwise stated, GeoRecon and Coord use the same pretraining data, backbone family, and 400k-step schedule, with GeoRecon adding only the clean-alignment and graph-conditioned reconstruction branches during pretraining. We first measure representation stability, then evaluate downstream performance on QM9, MD17, and MD22, and finally isolate representation quality with frozen-encoder linear probing.

5.1. Lipschitz Constant Analysis after Pretraining

We measured the Lipschitz constant of GeoRecon after pretraining on PCQM4Mv2 and observed a substantial reduction ($\sim 99\%$) in GeoRecon compared to Coord.

As shown in Table 1, GeoRecon consistently achieves median $L(x) \approx 30$ with negligible variance across step sizes, while the Coord baseline yields values on the order of $10^4 \sim 10^5$, representing a reduction of roughly *three orders of magnitude* in the local Lipschitz constant.

Figure 2 compares the sensitivity of GeoRecon and Coord encoders near equilibrium conformations. Under small perturbations ($\|\delta\boldsymbol{x}\| \leq 0.4$ Å), GeoRecon exhibits much smoother representation manifolds, with smaller changes in $\|\boldsymbol{h}\|$ compared to Coord. This trend is consistent both in individual samples and in averages across multiple molecules.

The experimental results confirm our intuition that anchoring a graph-level perspective during pretraining leads to a smoother latent space. This provides a stronger theoretical foundation for our improvements in downstream tasks, as smoother representation spaces have been shown to facilitate better learning procedures in various other settings (Lee et al., 2025; Guo et al., 2024; Zhang et al., 2025; Krishnan et al., 2020).

5.2. Test on Downstream Tasks

Full device, hyperparameter, and cost details are provided in Appendix N.

Given the reduced local sensitivity observed above, we expect GeoRecon to yield more robust features and improved generalization when fine-tuned for downstream prediction. Our method is built upon coordinate denoising with an additional graph-conditioned reconstruction objective. Therefore, *the primary controlled comparison is the denoising baseline* ***Coord*** (Zaidi et al., 2023), where data, backbone, and training schedule are matched. To highlight improvements over this direct baseline, we mark them in green in Tables 2, 3, and 4.

We also include non-pretrained and pretrained baselines to broaden empirical coverage. The graph-level and multimodal pretraining baselines, including MoleculeSDE, are discussed in Appendix M; this makes clear how they relate to GeoRecon's graph-level objective while keeping Coord as the primary baseline for isolating the effect of graph-conditioned reconstruction. Additional appendix results evaluate larger recent datasets and provide controlled ablations isolating noise scale, pooling, decoder capacity, and clean/reconstruction loss terms.

Table 2. MAE (↓) on QM9 property prediction. Best and second-best results are **bolded** and underlined, respectively. Aver-rank is the mean of per-metric ranks, with ties averaged. The methods are divided into two groups: training from scratch and pretraining then finetuning. Green text indicates the relative improvement (%) over the base model; red indicates degradation.

Task *Unit*	μ (D)	α (a_0^3)	homo (meV)	lumo (meV)	gap (meV)	R^2 (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{\text{cal}}{\text{mol·K}}$)	Average rank
SE(3)-Trans	0.052	0.143	36.0	36.0	59.0	1.969	5.517	74.00	68.00	72.00	68.00	0.068	12.75
MoleculeSDE (VE)	0.027	0.056	25.8	21.63	41.84	0.233	1.474	10.95	11.04	10.71	11.47	0.029	9.17
SchNet	0.033	0.235	41.0	34.0	63.0	0.070	1.70	14.00	19.00	14.00	14.00	0.033	11.21
EGNN	0.029	0.071	29.0	25.0	48.0	0.106	1.55	11.00	12.00	12.00	12.00	0.031	10.25
DimeNet++	0.030	0.044	24.6	19.5	32.6	0.330	1.21	6.32	6.28	6.53	7.56	0.023	5.71
PaiNN	0.012	0.045	27.6	20.4	45.7	0.070	1.28	5.85	5.83	5.98	7.35	0.024	4.63
SphereNet	0.025	0.045	22.8	18.9	31.1	0.270	**1.12**	6.26	6.36	6.33	7.78	**0.022**	4.88
TorchMD-Net	**0.011**	0.059	20.3	17.5	36.1	**0.033**	1.84	6.15	6.38	6.16	7.62	0.026	5.21
Transformer-M	0.037	**0.041**	17.5	16.2	**27.4**	0.075	1.18	9.37	9.41	9.39	9.63	**0.022**	5.21
SE(3)-DDM	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024	6.38
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026	6.88
Coord	0.016	0.052	17.7	14.7	31.8	0.450	1.71	6.57	6.11	6.45	6.91	0.027	5.75
GeoRecon (Ours)	0.013	0.045	**16.3**	**14.0**	30.1	0.160	1.48	**5.15**	**5.09**	**5.13**	**6.39**	0.022	**3.00**
Relative Gain	18.8%	13.5%	7.9%	4.8%	5.3%	64.4%	13.5%	21.6%	16.7%	20.5%	7.5%	18.5%	

Table 3. MAE (↓) of force prediction on the MD17 dataset (kcal/mol/Å). Published baseline rows are taken from their original papers for context; Coord/GeoRecon and AniDS/AniDS w/ GeoRecon are our matched runs. Within each matched pair, the lower value is **bolded**.

Method	Aspirin	Benzene	Ethanol	Malonaldehyde	Naphthalene	Salicylic Acid	Toluene	Uracil
Coord	0.23299	0.15052	0.10909	0.16218	0.06266	0.13336	0.06843	0.09109
GeoRecon (Ours)	**0.21097**	**0.14734**	**0.09771**	**0.15705**	**0.05755**	**0.11996**	**0.06064**	**0.08716**
Rel. Gain (Coord)	9.45%	2.11%	10.43%	3.16%	8.14%	10.05%	11.37%	4.32%
Frad (Feng et al., 2023)	0.20870	0.19940	0.09100	0.14150	0.05300	0.10810	0.05400	0.07600
SliDe (Ni et al., 2024)	0.17400	0.16910	0.08820	0.15380	0.04830	0.10060	0.05400	0.08250
AniDS (Liu et al., 2025a)	0.12935	0.13937	0.05938	0.09860	0.04357	0.08458	0.04575	0.07565
AniDS w/ GeoRecon	**0.12878**	**0.12945**	**0.05222**	**0.09438**	**0.04255**	**0.08355**	**0.04562**	**0.07301**

5.2.1. Evaluation on QM9

QM9 benchmark (Ramakrishnan et al., 2014) contains 134k stable small organic molecules (up to 9 heavy atoms: C, O, N, F, and H), each optimized at the B3LYP/6-31G(2df,p) level of DFT. It provides geometric, energetic, electronic, and thermodynamic properties; in this work, we focus on the 3D geometries and associated properties relevant for molecular representation learning.

We use QM9 to assess whether graph-conditioned reconstruction improves graph-level property prediction after finetuning. See Table 2 for results. Baseline details are provided in Appendix M. GeoRecon achieves strong performance across a range of molecular property prediction tasks and obtains the best results on several energy-related targets, including U, U_0, H, and G. Compared with Coord, GeoRecon improves all reported targets under the matched setting. Overall, the QM9 results support that graph-level conditioning can strengthen downstream property prediction beyond coordinate denoising alone, with non-uniform gains across targets.

5.2.2. Evaluation on MD17

MD17 benchmark (Chmiela et al., 2017) consists of *ab initio* molecular dynamics (MD) trajectories for small molecules (e.g., ethanol, malonaldehyde, glycine) at 500 K, providing tens of thousands of DFT-computed total energies and atomic forces per molecule. It has become a widely used benchmark for assessing machine-learned force fields in the low-dimensional, gas-phase regime.

As shown in Table 3, GeoRecon consistently improves Coord across all force prediction tasks. These results show graph-level reconstruction yields gains over coordinate denoising under matched training settings. The same reconstruction signal also improves AniDS, a more recent SE(3)-equivariant denoising framework, on all eight MD17 molecules. This suggests GeoRecon is not tied to Coord specifically and can complement stronger modern denoising pretraining frameworks.

5.2.3. Evaluation on MD22

MD22 benchmark (Chmiela et al., 2023) extends MD17 to larger and more complex supramolecular systems (42–370 atoms), including peptides, lipids, carbohydrates, DNA base

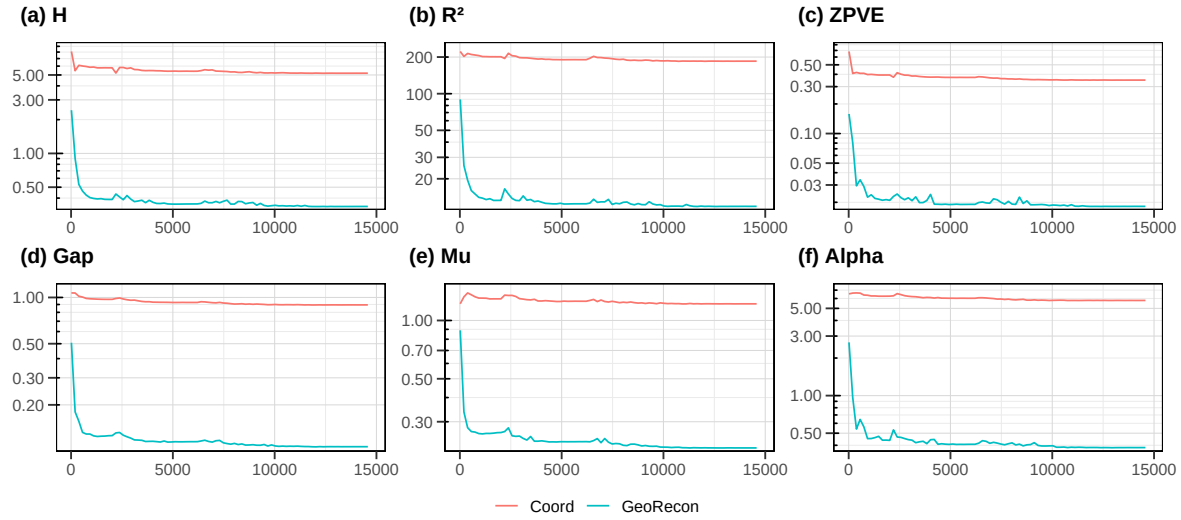

Figure 3. MAE (↓) loss curves of linear probing on multiple QM9 downstream tasks. The horizontal axis indicates the training step, and the vertical axis denotes the mean absolute error (MAE). The encoder is frozen, and only the linear output head is optimized. The training configuration is identical to that of the main experiments, ensuring a fair and controlled comparison. A cosine learning rate schedule with warmup is used, and all models are trained for 14.7k steps.

Table 4. MAE (↓) of different models on the MD22 dataset. Energy (kcal), Force (kcal/Å).

Molecule	DHA		Stachyose		AT-AT-CG-CG	
Metric	Energy	Force	Energy	Force	Energy	Force
Coord	0.15772	0.13891	0.70198	0.52737	0.48079	0.34858
GeoRecon	**0.11955**	**0.13553**	**0.59882**	**0.51666**	**0.47458**	**0.34080**
Relative Gain	24.20%	2.43%	14.70%	2.03%	1.29%	2.23%

pairs, and nanotube complexes. It provides *ab initio* MD trajectories at 400–500 K with a 1 fs timestep, and energies and forces computed at the PBE+MBD level of DFT, posing a greater challenge for ML-based force fields due to its size, flexibility, and strong non-local interactions.

We evaluate our method on three MD22 molecules: DHA (56 atoms), Stachyose (87 atoms), and AT-AT-CG-CG (118 atoms), covering small-, medium-, and large-scale systems. As shown in Table 4, GeoRecon outperforms the Coord backbone on both energy and force prediction across all three molecules, demonstrating its effectiveness beyond small-molecule scenarios.

5.3. Linear Probing Experiments

To isolate representation quality from finetuning dynamics, we conduct linear probing with a frozen encoder and train only a lightweight prediction head. Prior work (Kumar et al., 2022) shows that linear probing can preserve pretrained features and improve robustness in some out-of-distribution settings. Motivated by this observation, we compare GeoRecon and Coord under the same frozen-encoder protocol. Because the encoder is fixed, the linear head has limited capacity, so performance depends strongly on pretrained representation quality. We follow the same training configuration as the main experiments and train the linear head for 14.7k steps with a cosine learning rate schedule and warmup.

The results are shown in Figure 3. GeoRecon consistently outperforms Coord across all tasks under linear probing. Coord exhibits limited optimization in this setting, suggesting that its pretrained features require full finetuning to adapt effectively to downstream tasks. This observation is consistent with the Lipschitz measurements in the Introduction and indicates that GeoRecon learns smoother and more readily usable representations.

6. Conclusion

GeoRecon is a self-supervised, model-agnostic pretraining framework that uses only 3D coordinates to bridge coordinate denoising with graph-level supervision for molecular representation learning.

The core design is to condition reconstruction on a pooled graph embedding $\boldsymbol{g}$, broadcast to all atoms, together with a lightweight decoder and stronger reconstruction noise. This avoids the orientation ambiguity of decoding absolute coordinates from an invariant graph embedding, reduces trivial per-node shortcuts, and encourages $\boldsymbol{g}$ to carry molecule-wide information while remaining useful for atom-wise prediction.

As a mechanism, this graph-anchored conditioning yields a smoother and less sensitive representation space. Empirically, GeoRecon reduces local Lipschitz constants by nearly

three orders of magnitude compared with node-centric baselines, which aligns with the intuition that smoother representations support more robust downstream learning under structural perturbations.

Across QM9, MD17, and MD22, GeoRecon achieves competitive results and consistently improves upon the direct denoising baseline Coord. Linear probing further supports that the gains come from higher-quality pretrained representations. Our current limitations are also clear: improvements are uneven across all QM9 targets, pretraining is more expensive than Coord, and broader same-protocol comparisons with recent graph-level methods remain valuable. The appendix reports additional dataset results, ablations, and failure-case analyses to make these trade-offs explicit.

References

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Blum, L. C. and Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131(25):8732–8733, 2009. doi: 10.1021/ja902302h.

Chen, R., Li, C., Wang, L., Liu, M., Chen, S., Yang, J., and Zeng, X. Pretraining graph transformer for molecular representation with fusion of multimodal information. *Information Fusion*, 115:102784, 2025. doi: 10.1016/j.inffus.2024.102784.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. doi: 10.1126/sciadv.1603015.

Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., and Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9 (2):eadf0873, 2023. doi: 10.1126/sciadv.adf0873.

Feng, S., Ni, Y., Lan, Y., Ma, Z.-M., and Ma, W.-Y. Fractional denoising for 3D molecular pre-training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9938–9961. PMLR, 2023. URL https://proceedings.mlr.press/v202/feng23c.html.

Fu, C., Lin, Y., Krueger, Z., Yu, W., Qian, X., Yoon, B.-J., Arróyave, R., Qian, X., Maeda, T., Nakata, M., and Ji, S. A benchmark for quantum chemistry relaxations via machine learning interatomic potentials. *arXiv preprint arXiv:2506.23008*, 2025.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020.

Gasteiger, J., Giri, S., Margraf, J. T., and Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020a.

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=B1eWbxStPH.

Gasteiger, J., Becker, F., and Günnemann, S. GemNet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6790–6802, 2021.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021. doi: 10.1007/s10994-020-05929-w.

Guo, J., Xu, X., Pu, Y., Ni, Z., Wang, C., Vasu, M., Song, S., Huang, G., and Shi, H. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7548–7558, 2024. doi: 10.1109/CVPR52733.2024.00721.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022. doi: 10.1145/3534678.3539321.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Hussain, M. S., Zaki, M. J., and Subramanian, D. Triplet interaction improves graph transformers: Accurate molecular graph learning with triplet graph transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20768–20792. PMLR, 2024. URL https://proceedings.mlr.press/v235/hussain24a.html.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ji, X., Wang, Z., Gao, Z., Zheng, H., Zhang, L., Ke, G., and E, W. Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1489.

Jiao, R., Han, J., Huang, W., Rong, Y., and Liu, Y. Energy-motivated equivariant pretraining for 3D molecular graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8096–8104, 2023. doi: 10.1609/aaai.v37i7.25978.

Khrabrov, K., Ber, A., Tsypin, A., Ushenin, K., Rumiantsev, E., Telepov, A., Protasov, D., Shenbin, I., Alekseev, A., Shirokikh, M., Nikolenko, S., Tutubalina, E., and Kadurin, A. ∇^2DFT: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36869–36889, 2024. doi: 10.52202/079017-1162.

Krishnan, V., Al Makdah, A. A., and Pasqualetti, F. Lipschitz bounds and provably robust training by laplacian smoothing. *Advances in Neural Information Processing Systems*, 33:10924–10935, 2020.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Lee, H., Kim, M., Jang, S., Jeong, J., and Hwang, S. J. Enhancing variational autoencoders with smooth robust latent encoding. *arXiv preprint arXiv:2504.17219*, 2025.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., and Song, S. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv preprint arXiv:2012.11175*, 2020.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

Liu, S., Du, W., Ma, Z.-M., Guo, H., and Tang, J. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21497–21526. PMLR, 2023a. URL https://proceedings.mlr.press/v202/liu23h.html.

Liu, S., Guo, H., and Tang, J. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=CjTHVo1dvR.

Liu, X., Jiao, R., Liu, Z., Liu, Y., Liu, Y., Lu, Z., Huang, W., Zhang, Y., and Cao, Y. Learning 3D anisotropic noise distributions improves molecular force fields. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=tMBPJureSx.

Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B., and Ji, S. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021.

Liu, Y., Chen, J., Jiao, R., Li, J., Huang, W., and Su, B. DenoiseVAE: Learning molecule-adaptive noise distributions for denoising-based 3D molecular pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=ym7pr83XQr.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Luo, S., Chen, T., Xu, Y., Zheng, S., Liu, T.-Y., Wang, L., and He, D. One transformer can understand both 2D & 3D molecular data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vZTp1oPV3PC.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mobley, D. L. and Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014. doi: 10.1007/s10822-014-9747-x.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023. doi: 10.1038/s41467-023-36329-y.

Ni, Y., Feng, S., Ma, W.-Y., Ma, Z.-M., and Lan, Y. Sliced denoising: A physics-informed molecular pretraining method. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=liKkG1zcWq.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., Wang, K., and Tang, J. GCC: Graph contrastive coding for graph neural network pre-training. In *Proceedings*

of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1150–1160, 2020. doi: 10.1145/3394486.3403168.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

Ramakrishnan, R., Hartmann, M., Tapavicza, E., and von Lilienfeld, O. A. Electronic spectra from tddft and machine learning in chemical space. *The Journal of chemical physics*, 143(8):084111, 2015. doi: 10.1063/1.4928757.

Ren, G.-P., Yin, Y.-J., Wu, K.-J., and He, Y. Force field-inspired molecular representation learning for property prediction. *Journal of Cheminformatics*, 15(1):17, 2023. doi: 10.1186/s13321-023-00691-2.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12559–12571, 2020.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012. doi: 10.1103/PhysRevLett.108.058301.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. SchNet: A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Schütt, K. T., Unke, O. T., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388. PMLR, 2021. URL https://proceedings.mlr.press/v139/schutt21a.html.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. 3D infomax improves GNNs for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502. PMLR, 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Thölke, P. and De Fabritiis, G. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zNHzqZ9wrRB.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. doi: 10.1162/neco_a_00142.

Wang, L., Liu, S., Rong, Y., Zhao, D., Liu, Q., Wu, S., and Wang, L. MolSpectra: Pre-training 3D molecular representation with multi-modal energy spectra. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xJDxVDG3x2.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018. doi: 10.1039/c7sc02664a.

Xia, J., Zhao, C., Hu, B., Gao, Z., Tan, C., Liu, Y., Li, S., and Li, S. Z. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jevY-DtiZTR.

Yang, K., Swanson, K., Jin, W., Coley, C. W., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B. P., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. F., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P. W., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tYIMtogyee.

Zhang, F., Fan, X., Su, X., and Gao, G. Repcali: High efficient fine-tuning via representation calibration in latent space for pre-trained language models. *arXiv preprint arXiv:2505.08463*, 2025.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15870–15882, 2021.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. 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.

Zhu, J., Xia, Y., Wu, L., Xie, S., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2626–2636, 2022. doi: 10.1145/3534678.3539368.

A. Impact Statement

This work aims to advance machine learning for 3D molecular representation learning. Potential benefits include accelerating chemistry and materials workflows. Potential risks include dual-use in harmful chemical design, dataset bias in downstream deployment, and additional pretraining cost. These risks should be addressed through domain oversight, careful evaluation, and transparent reporting.

B. Extended Related Work

Geometric Graph Neural Networks Early methods incorporated 3D geometry into GNNs via continuous-filter convolutions and directional message passing. SchNet (Schütt et al., 2018) operates directly on Cartesian coordinates to predict energy and forces. DimeNet (Gasteiger et al., 2020b) and DimeNet++ (Gasteiger et al., 2020a) capture angular information, improving performance and efficiency. GemNet (Gasteiger et al., 2021) further enhances accuracy using multi-hop and higher-order geometric features. Meanwhile, EGNN (Satorras et al., 2021) and SE(3)-Transformer (Fuchs et al., 2020) encode Euclidean symmetries through equivariance, matching tensor-based models with lower complexity.

Implicit Use of 3D Information Some prior works implicitly leverage 3D structural information by converting 1D SMILES strings or 2D molecular graphs into 3D conformations using cheminformatics tools such as RDKit (Liu et al., 2022; Zhu et al., 2022; Stärk et al., 2022). These derived conformers are then used during training to enhance model capacity, achieving strong results on benchmarks like PCQM4Mv2 (Hussain et al., 2024), which are originally designed for 2D graph modalities.

Self-Supervised Denoising Pretraining Self-supervised denoising has emerged as a principled means to leverage unlabeled molecular conformations. Coord pretraining (Zaidi et al., 2023) demonstrates that training a GNN to remove Gaussian perturbations from equilibrium structures is mathematically equivalent to learning the underlying molecular force field, markedly improving downstream QM9 performance under limited labels. Frad (Feng et al., 2023) furthers this concept by combining dihedral-angle and Cartesian perturbations, then denoising only the coordinate fraction to preserve force-learning equivalence while broadening low-energy conformer coverage. SliDe (Ni et al., 2024) perturbs bond lengths, angles, and torsions under classical intramolecular potential theory and uses random slicing to avoid expensive Jacobian computation. DenoiseVAE (Liu et al., 2025b) instead learns molecule- and atom-adaptive noise distributions with a VAE-style noise generator, addressing the limitation of shared hand-designed noise schedules. AniDS (Liu et al., 2025a) further moves from isotropic or homoscedastic noise to structure-aware anisotropic Gaussian covariances that are atom-specific and SO(3)-equivariant. More recently, MolSpectra (Wang et al., 2025) integrates multimodal energy-level spectra via contrastive reconstruction, enriching the model's access to discrete quantum information without external labels.

Multi-Task and Multi-Fidelity Learning To address data scarcity in quantum chemistry, one strategy is to exploit shared structure across tasks or fidelity levels. In *multi-task learning*, models predict several quantum properties simultaneously, e.g., 12 QM9 targets, encouraging a shared embedding space and reducing overfitting. In *multi-fidelity learning*, one leverages the hierarchy of computational costs: accurate methods like CCSD(T) are expensive, while cheaper ones like DFT scale to larger datasets. Recent work (Ren et al., 2023) combines abundant DFT data with limited CC samples to approach CC-level accuracy at far lower cost. Together, these strategies mitigate scarcity by balancing accuracy and efficiency.

Graph-Level Pretraining in Molecular Representation Learning Early GNN pretraining mainly targeted node-level tasks such as masked atom prediction (Hu et al., 2020; Hou et al., 2022; Xia et al., 2023), but recent efforts increasingly adopt graph-level objectives to capture molecule-wide signals. GROVER (Rong et al., 2020) introduces motif prediction to identify recurring substructures, and MGSSL (Zhang et al., 2021) extends this by generating graphs motif-by-motif, injecting global inductive biases. GraphCL (You et al., 2020) and GCC (Qiu et al., 2020) apply graph-level contrastive learning with augmented graph views to enforce whole-graph consistency.

Other approaches leverage cross-modal or multi-scale alignment: GraphMVP (Liu et al., 2022) aligns 2D graphs with 3D conformers via contrastive learning, MoleculeSDE (Liu et al., 2023a) uses group-symmetric SDEs to connect 2D and 3D modalities, and Uni-Mol2 (Ji et al., 2024) uses a two-track transformer jointly encoding atomic, graph, and spatial features. MolGT (Chen et al., 2025) combines node- and graph-level pretext tasks across 2D topology and 3D geometry. Though effective, these methods often require additional modalities, dedicated augmentation pipelines, or more complex training designs.

3D-EMGP (Jiao et al., 2023) is especially relevant because it also combines node-level geometry supervision with a graph-level pretraining signal. Its graph-level component predicts the noise scale in an energy-motivated equivariant pretraining framework. GeoRecon differs in the role assigned to the graph representation: rather than predicting a graph-level scalar, GeoRecon broadcasts a pooled graph embedding into a weak reconstruction-as-denoising decoder, so the graph embedding must support atom-wise recovery un-

der stronger corruption. This keeps the framework close to coordinate denoising while making graph-level conditioning the mechanism of supervision.

Positioning and Gaps Existing molecular pretraining methods differ along several key dimensions that expose a gap in how global structure is enforced. First, many strong graph-level objectives rely on 2D graphs, cross-modal alignment, or additional signals beyond 3D coordinates, e.g., 2D–3D contrastive alignment (Liu et al., 2022) or multi-modal spectra (Wang et al., 2025), whereas we target a *3D-only* setting. Second, objectives span masked/attribute prediction and contrastive learning (Hu et al., 2020; You et al., 2020; Qiu et al., 2020) to node-level denoising (Zaidi et al., 2023; Feng et al., 2023; Ni et al., 2024), but these typically optimize *local* supervision that does not explicitly demand molecule-wide coherence. Third, even when graph-level signals are present, they are often introduced via multi-view augmentations, multi-task formulations, or separate graph-level prediction heads. We focus instead on *graph-level conditioning*: by broadcasting a pooled graph embedding into a reconstruction-as-denoising head, GeoRecon injects global supervision while keeping the backbone and training recipe close to standard coordinate denoising. These observations motivate our design of GeoRecon.

C. Ablation Studies

C.1. Ablation of Task Rec and Clean

To assess the contributions of different hyperparameters in our GeoRecon framework, we conduct ablation studies along three axes: decoder depth L, reconstruction noise scale λ, and reconstruction loss weight $\lambda_{\texttt{rec}}$.

Effect of Reconstruction Noise Scale λ. In subsection 4.1 we pointed out the influence of the reconstruction noise scale λ, which directly modulates the difficulty of the pretraining task. To further validate our conclusions, we performed additional experiments. Table 5 reports both the measured Lipschitz constants of the learned representations and the downstream performance on the QM9 H_{atom} task, using a 14-layer encoder pretrained on a 10k subset of QM9 with varying λ.

Table 5. Effect of reconstruction noise scale λ on Lipschitz constant (Lip) and downstream MAE (↓) for the H_{atom} task. Pretraining is performed on a 10k subset of QM9 with a 14-layer encoder.

λ	1.15	1.20	1.35	1.50
Lip	64.505	59.934	59.489	62.945
H_{atom}	7.0172	6.5091	**6.4686**	6.6525

Additional controlled comparisons on PCQM4Mv2 (100k subset). To compare against stronger denoising baselines and better understand the effect of λ at larger pretraining scale, we ran additional controlled experiments using the TorchMD-Net backbone. We pretrained on a 100k subset of PCQM4Mv2 for 400k steps and then finetuned on QM9. All reported numbers below are MAE (↓).

Table 6. Comparison to the strong node-level baseline Frad on the QM9 R2 target (electronic spatial extent). GeoRecon variants are denoted by their clean and reconstruction loss weights (Clean$\lambda_{\texttt{cln}}$/Rec$\lambda_{\texttt{rec}}$).

	Frad	Clean1/Rec0.5	Clean5/Rec0.5	Clean10/Rec0.5	Clean10/Rec0.35	Clean10/Rec0.2	Clean10/Rec0.05
QM9 R2	0.417721	**0.258833**	0.285035	0.482544	0.293071	<u>0.259997</u>	0.263396

Table 7. Expanded reconstruction noise-scale sweep on PCQM4Mv2 (100k subset) with finetuning on QM9.

λ	1.00	1.05	1.10	1.15	1.20	1.35	1.50
R2	0.4244	**0.2914**	0.3466	0.6667	0.3818	<u>0.3327</u>	0.5148
α	0.0771	0.0656	<u>0.0616</u>	0.0727	0.0675	**0.0589**	0.0625
H	6.5887	6.9864	6.8120	7.0172	<u>6.5091</u>	**6.4686**	6.6525
μ	0.0164	<u>0.0156</u>	0.0171	0.0172	**0.0152**	0.0160	<u>0.0156</u>

Table 8. Full QM9 evaluation for reconstruction noise-scale variants (MAE ↓).

Method	μ	α	HOMO	LUMO	Gap	R^2	ZPVE	C_v
Coord	0.01600	0.05200	0.01770	0.01470	0.03180	0.4500	1.710	0.02700
GeoRecon ($\lambda = 1.00$)	**0.01200**	0.04800	0.01690	0.01420	0.03140	0.2380	1.510	**0.02200**
GeoRecon ($\lambda = 1.35$)	0.01306	**0.04479**	**0.01628**	**0.01402**	**0.03006**	**0.1598**	**1.481**	0.02398

As shown, moderate values of λ (e.g., 1.20–1.35) yield both reduced Lipschitz constants and improved downstream accuracy, suggesting smoother and more transferable representations. This observation aligns with our intuition: larger λ injects stronger noise, making reconstruction more challenging and forcing the model to capture global molecular dependencies, whereas excessively large λ degrades stability. Notably, the λ values used in our main experiments were not exhaustively tuned, implying that additional calibration may further enhance GeoRecon's performance.

Stronger noise alone is insufficient on MD17. To isolate the effect of graph-level reconstruction from the effect of simply increasing denoising difficulty, we compare Coord with noise scale 1.0, Coord with the stronger reconstruction noise scale 1.35, and GeoRecon with the same 1.35 scale. All models use matched downstream settings.

Table 9 shows that stronger perturbation alone is not a reliable explanation for the gains. It slightly helps Coord on ethanol and naphthalene but hurts on malonaldehyde and salicylic acid, while graph-conditioned reconstruction improves all four molecules.

Effect of Decoder Depth and Noise Scale. We vary the number of layers $L \in \{3, 4, 5\}$ in the lightweight decoder and the reconstruction noise scale $\lambda \in \{1.0, 1.5\}$, which

Table 9. MD17 force MAE (↓) under stronger noise-scale controls. Increasing denoising noise alone does not consistently improve Coord, whereas GeoRecon improves under the same scale.

Molecule	Coord ($\lambda = 1.0$)	Coord ($\lambda = 1.35$)	GeoRecon ($\lambda = 1.35$)
Ethanol	0.10909	0.10722	**0.09771**
Malonaldehyde	0.16218	0.16396	**0.15705**
Naphthalene	0.06266	0.06126	**0.05755**
Salicylic Acid	0.13336	0.18087	**0.11996**

controls the magnitude of perturbation in the reconstruction targets.

Table 10. Effect of decoder depth L and noise scale λ on finetuning performance (MAE↓). Each group reports HOMO, LUMO, and GAP errors; the best result in each task is highlighted in **bold**, and the second-best is underlined. (10 pretraining epochs)

λ	Task	$L = 3$	$L = 4$	$L = 5$
1.0	HOMO	0.01812	0.02028	0.01687
	LUMO	0.01506	0.02110	**0.01368**
	GAP	**0.03132**	0.04220	0.03157
1.5	HOMO	0.01722	**0.01661**	0.01688
	LUMO	0.01472	0.01451	0.01497
	GAP	0.03218	0.03307	0.03238

The results are shown in Table 10. They suggest an interaction between decoder depth and noise scale. When the lightweight decoder is shallow (e.g., $L = 3$ or $L = 4$), increasing the noise scale from $\lambda = 1.0$ to $\lambda = 1.5$ improves downstream performance, likely because stronger perturbations force the encoder to rely more heavily on the global representation to support accurate reconstruction.

However, when the decoder becomes deeper ($L = 5$), it gains sufficient capacity to locally denoise without depending as much on the global context. As a result, increasing the noise scale may not bring further benefit, and can even lead to performance degradation due to a mismatch between input difficulty and the model's reliance on global structure. The results suggest that shallower decoders rely more on graph-level guidance, supporting our core hypothesis about the need for global conditioning.

Effect of Reconstruction Loss Weight. In our ablation study with 30-epoch pretraining and a CosineWarmup schedule, we varied the reconstruction loss weight $\lambda_{\texttt{rec}} \in \{0.40, 0.45, 0.50\}$ while fixing decoder depth $L = 5$ and noise scale $\lambda = 1.0$.

The results in Table 11 reveal a non-monotonic trend as $\lambda_{\texttt{rec}}$ varies. A low weight (0.40) leads to better performance on HOMO, while higher values (e.g., 0.50) improve LUMO and gap. The intermediate setting ($\lambda_{\texttt{rec}} = 0.45$) achieves the most balanced performance. The results suggest that $\lambda_{\texttt{rec}}$ modulates the encoder's representational focus. We interpret this as an empirical observation that stronger reconstruction constraints may influence the encoder's global focus, motivating further analysis.

Table 11. Effect of reconstruction loss weight $\lambda_{\texttt{rec}}$ with fixed $L = 5$, $\lambda = 1$, the best result in each task is highlighted in **bold**. (Pre-training 30 epoch)

$\lambda_{\texttt{rec}}$	HOMO	LUMO	GAP
0.4	**0.01630**	0.01509	0.03345
0.45	0.01756	**0.01379**	**0.03141**
0.5	0.01659	0.01403	0.03227

Pooling strategy. We also ablate the aggregation function used to form the graph-level conditioning vector. In this controlled run, mean pooling and attention pooling both substantially improve over the no-reconstruction baseline, while max pooling performs worse, consistent with the intuition that a single atom-level feature is too narrow to support molecule-wide reconstruction.

Table 12. Pooling-strategy ablation under noise scale $\lambda = 1.35$ after 300k finetuning steps. Lower MAE is better.

Variant	no-rec	Attn pool	Mean pool	Max pool
MAE	0.62482	0.37483	**0.36566**	0.57874

D. Analysis of Local Lipschitz Constants

We measured the spectral norm of the encoder Jacobian on PCQM4Mv2 samples for both GeoRecon and Coord after the same pretraining period, and define the *Local Lipschitz Constant* (LLC) at a conformation x as follows (Miyato et al., 2018):

$$L(x) = \|J_f(x)P\|_2$$

Here, f denotes the encoder, and $J_f(x) \in \mathbb{R}^{d\times 3N}$ is the Jacobian of f at x. The orthogonal projector $P \in \mathbb{R}^{3N\times 3N}$ onto the non-rigid subspace is applied to remove rigid-body degrees of freedom, so that the measured local Lipschitz constant reflects sensitivity only to physically meaningful (non-rigid) deformations and is independent of global SE(3) motions. Following Gouk et al. (2021) and Novak et al. (2018), we approximate $L(x)$ via power iteration.

E. Why Smoother Representations Improve Finetuning

Setting. Let $(\mathcal{X}, \bar{d})$ denote the space of molecular conformations modulo rigid motions, where $\bar{d}$ is the Procrustes / RMSD distance

$$\bar{d}(x, x') := \min_{R\in \mathrm{SO}(3),\, t\in\mathbb{R}^3} \big\|x - (Rx' + t)\big\|_2.$$

Let $f : (\mathcal{X}, \bar{d}) \to (\mathbb{R}^m, \|\cdot\|_2)$ be a *graph-level* representation map with Lipschitz constant L_f, i.e., $\|f(x) -$

$f(x')\|_2 \le L_f\, \bar{d}(x, x')$. We assume a training sample $S = \{(x_i, y_i)\}_{i=1}^n$ with diameter $\mathrm{diam}_{\bar{d}}(S) \le D$. The downstream loss ℓ is assumed to be 1-Lipschitz in its prediction argument and bounded (or rescaled to $[0,1]$ for concentration).[1] All results are stated on the SE(3)-*quotient* space and thus respect molecular symmetries.

E.1. Linear-Probe Analysis (Frozen Encoder)

We consider a linear probe on top of the frozen f: $g_w(z) = \langle w, z\rangle + b$ with $\|w\|_2 \le B$; denote $\mathcal{H}_{\mathrm{lin}} = \{x \mapsto g_w(f(x))\}$.

Theorem E.1 (Smoother representations tighten generalization for linear probes)**.** *With probability at least $1-\delta$, every $h \in \mathcal{H}_{\mathrm{lin}}$ satisfies*

$$\mathcal{R}(h) \;\le\; \widehat{\mathcal{R}}_S(h) \;+\; \frac{c\,B\,L_f\,D}{\sqrt{n}} \;+\; \sqrt{\frac{\log(1/\delta)}{2n}},$$

for an absolute constant c.

Sketch. Apply the contraction lemma to $\phi_y(u) = \ell(y,u) - \ell(y,0)$, which is 1-Lipschitz and satisfies $\phi_y(0) = 0$, to obtain $\mathfrak{R}_n(\ell \circ \mathcal{H}_{\mathrm{lin}}) \le \mathfrak{R}_n(\mathcal{H}_{\mathrm{lin}})$. For $\mathcal{H}_{\mathrm{lin}}$ we have

$$\mathfrak{R}_n(\mathcal{H}_{\mathrm{lin}}) \le \frac{B}{n}\,\mathbb{E}_\sigma\Big\|\sum_{i=1}^n \sigma_i f(x_i)\Big\|_2.$$

Pick any reference x_0 and set $c = f(x_0)$; the bias b absorbs this shift. Since $\|f(x_i) - c\|_2 \le L_f\, \bar{d}(x_i, x_0) \le L_f D$ for all i,

$$\begin{aligned}\mathbb{E}_\sigma\Big\|\sum_{i=1}^n \sigma_i\big(f(x_i) - c\big)\Big\|_2 &\le \Big(\sum_{i=1}^n \|f(x_i) - c\|_2^2\Big)^{1/2} \\ &\le \sqrt{n}\, L_f D.\end{aligned}$$

Hence $\mathfrak{R}_n(\mathcal{H}_{\mathrm{lin}}) \le (BL_fD)/\sqrt{n}$. Plugging this into the standard Rademacher generalization bound for bounded 1-Lipschitz losses yields (E.1). □

Lemma E.2 (Noise robustness scaling with L_f under a linear probe)**.** *Let the input be perturbed by additive Gaussian noise $x \mapsto x + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ in Cartesian coordinates (for N atoms hence $3N$ dimensions). Then for any w with $\|w\|_2 \le B$,*

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\big|g_w(f(x+\varepsilon)) - g_w(f(x))\big|\big] &\le B\,L_f\,\mathbb{E}_\varepsilon\|\varepsilon\|_2 \\ &= \Theta\big(B\,L_f\,\sigma\sqrt{3N}\big), \\ \mathbb{E}_\varepsilon\Big[\big(g_w(f(x+\varepsilon)) - g_w(f(x))\big)^2\Big] &\le B^2 L_f^2\,\mathbb{E}_\varepsilon\|\varepsilon\|_2^2 \\ &= 3N\sigma^2 B^2 L_f^2.\end{aligned}$$

[1] Boundedness is only used for the empirical-to-population concentration term.

Sketch. Alignment only decreases distance:

$$\begin{aligned}\bar{d}(x, x+\varepsilon) &= \min_{R,t} \|x - (R(x+\varepsilon) + t)\|_2 \\ &\le \|x - (x+\varepsilon)\|_2 = \|\varepsilon\|_2.\end{aligned}$$

Thus,

$$\begin{aligned}&|g_w(f(x+\varepsilon)) - g_w(f(x))| \\ &\quad\le \|w\|_2\,\|f(x+\varepsilon) - f(x)\|_2 \\ &\quad\le BL_f\bar{d}(x, x+\varepsilon) \le BL_f\|\varepsilon\|_2.\end{aligned}$$

Taking (squared) expectations with respect to ε gives the claims. □

The inequality $\bar{d}(x, x+\varepsilon) \le \|\varepsilon\|_2$ is a conservative upper bound, since the optimal alignment (R, t) typically yields a much smaller distance in practice. Therefore, our result should be interpreted as a worst-case guarantee: the amplification of input perturbations cannot exceed $BL_f\|\varepsilon\|_2$, while in realistic settings the effect is usually weaker.

Implication. Equations (E.1) and Lemma E.2 show that a smaller L_f (i.e., a smoother f) strictly tightens linear-probe generalization and attenuates noise amplification.

E.2. End-to-End Finetuning: Rigorous and Heuristic Routes

Consider now an end-to-end predictor $h = g \circ f$ where $g : \mathbb{R}^m \to \mathbb{R}$ is a small readout network (e.g., the MLPs used in our experiments). Let L_g denote the Lipschitz constant of g on the image of f over the data domain. The composite Lipschitz constant satisfies

$$L_h \;\le\; L_g L_f. \tag{1}$$

A rigorous route under spectral control.

Proposition E.3 (Generalization and robustness with spectrally-controlled readouts)**.** *Let g be a depth-d MLP with 1-Lipschitz activations (e.g., ReLU) and weight matrices $W_1, \dots, W_d$. Assume spectral norm bounds $\|W_j\|_{\mathrm{op}} \le s_j$ and define the product of spectral norms $P := \prod_{j=1}^d s_j$ and the complexity factor $S := \big(\sum_{j=1}^d \|W_j\|_F^2/s_j^2\big)^{1/2}$. Let $L_g \le P$ and $h = g \circ f$. Then, with probability at least $1-\delta$,*

$$\begin{aligned}\mathcal{R}(h) &\le \widehat{\mathcal{R}}_S(h) \\ &\quad + \tilde{c}\frac{PSL_fD}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}, \\ \mathbb{E}_\varepsilon\big[(h(x+\varepsilon) - h(x))^2\big] &\le (L_gL_f)^2 \cdot 3N\sigma^2.\end{aligned}$$

for a universal constant $\tilde{c}$ (independent of dimensions and model parameters).

Sketch. By (1), h is L_h-Lipschitz with $L_h \le L_g L_f$; the noise bound follows as in Lemma E.2. For generalization, apply a vector (Gaussian) contraction inequality to the composed class:

$$\mathfrak{R}_n(\{g \circ f\}) \le C \left(\prod_{j=1}^{d} s_j \right) \frac{1}{n} \mathbb{E} \Big\| \sum_{i=1}^{n} \gamma_i f(x_i) \Big\| \le C \left(\prod_{j=1}^{d} s_j \right) \frac{L_f D}{\sqrt{n}},$$

and incorporate the standard spectral/Frobenius control via $S = \left(\sum_{j=1}^{d} \|W_j\|_F^2 / s_j^2 \right)^{1/2}$ to obtain the stated $\tilde{c} \frac{PSL_f D}{\sqrt{n}}$ (constants absorbed into $\tilde{c}$). □

Remark. In our experiments, we *did not* explicitly impose spectral norm constraints via dedicated regularization. However, common practices such as weight decay and initialization schemes tend to implicitly control the spectral norms of the weights, preventing L_g and the complexity factor S from becoming excessively large. Thus Proposition Theorem E.3 should be interpreted primarily as a conceptual tool, clarifying how representation smoothness (L_f) and readout complexity jointly govern generalization.

Beyond such constrained settings, we next discuss a heuristic justification based on early-stage linearization.

A heuristic route via early-stage linearization. While the exact Rademacher complexity of general MLPs depends on architectural details, early end-to-end training is well-approximated by a linearized model around the pretrained parameters (NTK-style local linearization). Under this viewpoint, the estimation term scales as $\tilde{\mathcal{O}}(L_g L_f D/\sqrt{n})$ with L_g the *local* Lipschitz constant near initialization, and the noise bound remains governed solely by the composite Lipschitz constant $(L_g L_f)$.

To empirically support our heuristic argument on early-stage linearization, we conduct a sanity-check experiment measuring both the cosine similarity between the predictions of the full model and its NTK approximation, and the alignment of parameter gradients across steps. As shown in Table 13, both metrics remain high (> 0.85) during the first 300 steps, confirming that finetuning dynamics are well-approximated by a linear regime in the early stage. Although the similarity gradually decreases later, this evidence substantiates our claim that smoother encoders (smaller L_f) can benefit optimization through improved linearized dynamics.

As expected, the similarity metrics are highest in the very first steps and gradually decrease, which aligns with the known phenomenon that NTK approximations are most accurate during the initial phase of training.

Table 13. Sanity check for linearization regime during early finetuning. We report the cosine similarity between predictions of the full model and its linearized NTK approximation, as well as the gradient alignment, measured as the cosine similarity between parameter gradients at consecutive steps. Higher values indicate stronger validity of the linearization heuristic.

Step Range	Cosine Sim. (Full vs. NTK)	Grad. Alignment
0–100	0.98	1.000
100–200	0.90	0.927
200–300	0.87	0.921

Practical relevance. Two considerations explain why full finetuning can match or surpass linear probes: (i) *Certificate existence:* the hypothesis class for full finetuning strictly contains that of linear probes (by taking g to be linear and freezing f), so Theorem E.1 remains attainable as a special case. (ii) *Local linearization:* in early training, dynamics are well-approximated by a linearized model around the pretrained parameters, effectively optimizing a linear head on a fixed f and thus retaining the same L_f dependence; with spectral/weight decay control, L_g (and hence L_h) stays finite. If end-to-end finetuning underperforms compared to linear probing, this often suggests limitations in the optimization procedure's ability to leverage the expanded hypothesis class, rather than deficiencies in the representation itself.

Conclusion. Both the linear-probe guarantee (Theorem E.1) and the spectrally-controlled end-to-end bound (Proposition E.3), together with our empirical ablations, indicate that pretraining strategies which produce *smoother* graph-level encoders (smaller L_f) can improve finetuning generalization and reduce sensitivity to coordinate perturbations. Practically, spectral or norm-based regularization on the readout complements smaller L_f, yielding tighter overall control via $L_h \le L_g L_f$.

Scope of analysis. Our theoretical results should be interpreted at three different levels of rigor. (i) Theorem E.1 provides a tight and formally complete guarantee for the case of *linear probing*, where the dependence on L_f can be isolated exactly. (ii) Proposition E.3 extends this guarantee to nonlinear readouts under explicit spectral norm and Frobenius norm control, which mathematically preserves the same L_f-dependence but requires additional architectural constraints. (iii) Beyond these constrained settings, inspired by Jacot et al. (2018), we provide a heuristic justification via early-stage linearization, suggesting that the benefits of smoothness L_f are likely to extend to practical end-to-end finetuning. While the fully unconstrained nonlinear case remains an open theoretical question, our experiments in main content and Table 5 consistently support the practical value of smoother representations across both linear and

nonlinear downstream models.

Limitation. Our guarantees are rigorous for linear probes; extensions to nonlinear readouts rely on either architectural constraints or heuristic arguments, and the fully unconstrained case remains open. Nonetheless, our experiments consistently indicate that smoother representations (L_f) improve performance across both linear and nonlinear settings.

F. Proof of equivalence between denoise and force field learning

We briefly revisit the theoretical grounding of denoising as an approximation to physical force learning, as established by prior work (Zaidi et al., 2023).

Let $\boldsymbol{x}_{i(j)}$ denote the position of atom i in molecule j, where each conformation $\boldsymbol{x}_i \in \mathbb{R}^{3N}$ is mean-centered to ensure $\sum_j \boldsymbol{x}_{i(j)} = \mathbf{0}$, thereby lying in a $(3N-3)$-dimensional subspace. Denote the potential energy function as $U(\boldsymbol{x})$, which maps geometric configurations to their corresponding energy levels. The negative gradient $-\nabla_{\boldsymbol{x}} U(\boldsymbol{x})$ corresponds to the force field and serves as the prediction target in force-learning tasks. The distribution of molecular conformations $\boldsymbol{x}$ is denoted as $p_{\text{phy}}(\boldsymbol{x})$, which follows the Boltzmann distribution and takes the form $A \exp\left(-U(\boldsymbol{x})/kT\right)$ where A is a normalization constant, k is the Boltzmann constant, and T is the temperature.

From a probabilistic perspective, the distribution of the noised coordinates $\boldsymbol{x}^{\texttt{nsd}}$ can be expressed as a marginalization over the clean coordinates:

$$p(\boldsymbol{x}^{\texttt{nsd}}) = \int p_\tau(\boldsymbol{x}^{\texttt{nsd}} | \boldsymbol{x}^{\texttt{cln}}) p(\boldsymbol{x}^{\texttt{cln}}) \, \mathrm{d}\boldsymbol{x}^{\texttt{cln}}$$

Here, the conditional distribution $p_\tau(\boldsymbol{x}^{\texttt{nsd}} \mid \boldsymbol{x}^{\texttt{cln}})$ is modeled as a Gaussian in the $(3N-3)$-dimensional mean-centered subspace, with isotropic variance τ^2. According to Vincent (2011), denoising score matching (DSM) is equivalent to explicit score matching (ESM). Writing $\boldsymbol{x}_n = \boldsymbol{x}^{\texttt{nsd}}$, $\boldsymbol{x}_c = \boldsymbol{x}^{\texttt{cln}}$, and $G_\theta = \text{GNN}_\theta$,

$$\begin{aligned}
\mathcal{L}_{\text{DSM}} &= \mathbb{E}_{p_\tau(\boldsymbol{x}_n, \boldsymbol{x}_c)} \left\| G_\theta(\boldsymbol{x}_n) - \frac{\partial \log p_\tau}{\partial \boldsymbol{x}_n} \right\|^2 \\
&= \mathbb{E}_{p_\tau(\boldsymbol{x}_n, \boldsymbol{x}_c)} \left\| G_\theta(\boldsymbol{x}_n) + \frac{1}{\tau^2}(\boldsymbol{x}_n - \boldsymbol{x}_c) \right\|^2 \\
&= \mathbb{E}_{p_{\text{phy}}(\boldsymbol{x}_n)} \left\| G_\theta(\boldsymbol{x}_n) - (-\nabla U(\boldsymbol{x}_n)) \right\|^2 + \text{Const} \\
&= \mathcal{L}_{\text{ESM}} + \text{Const}.
\end{aligned}$$

This establishes a theoretical bridge between denoising-based supervision and force field learning under equilibrium statistics. Accordingly, our reconstruction task, which leverages graph-level embeddings to modulate scaling denoising, is not only empirically effective but also theoretically justified through its approximation of physically meaningful gradients. This connection underlies our decision to instantiate reconstruction via a scaling denoise mechanism.

This interpretation is also supported by the general connection between denoising and score matching established in Vincent (2011).

G. Ablation Study on Clean and Reconstruction Tasks

To isolate the contributions of the reconstruction task and the clean alignment task, we conduct an additional ablation. The RecOnly model sets the clean-task weight to 0 while keeping other hyperparameters unchanged.

Table 14. Effect of clean alignment (MAE↓) with fixed $L = 5$, $\lambda = 1$, under three pretraining settings: Baseline, RecOnly, and Rec+Clean. The best result for each target is highlighted in **bold**. (30 pretraining epochs)

Model	HOMO	LUMO	GAP
Coord	0.0177	0.0147	0.0318
RecOnly	**0.0163**	0.0141	0.0315
Rec+Clean	0.0166	**0.0137**	**0.0306**

The results are shown in Table 14. Rec+Clean achieves the best performance on most targets. RecOnly also improves over Coord, while removing the clean task slightly degrades performance compared with Rec+Clean, suggesting that clean alignment contributes to the overall gain.

H. Evaluation on 3BPA

3BPA dataset (Musaelian et al., 2023) consists of *ab initio* MD trajectories for the flexible drug-like molecule 3-(benzyloxy)pyridin-2-amine (3BPA), characterized by three central rotatable dihedral angles that yield a complex potential energy surface with multiple local minima. Configurations were generated via MD simulations at 300, 600, and 1200 K (25 ps length, 1 fs timestep, Langevin thermostat) and re-evaluated using DFT (ωB97X/6-31G(d)). This dataset provides an explicit out-of-distribution (OOD) scenario absent in equilibrium datasets such as QM9 and MD17, making it a stringent testbed for generalization.

To assess the robustness of GeoRecon, we pretrained and finetuned both GeoRecon and its backbone model (TorchMD-Net) on 3BPA for 100 epochs under identical hyperparameters. Training was performed using conformations sampled at 300 K, while evaluation covered conformations at 300, 600, and 1200 K. Results are summarized in

Table 15.

Table 15. MAE (↓) on the 3BPA dataset at different temperatures.

Method	Metric	300K	600K	1200K
TorchMD-Net	Energy	0.08606	0.08734	0.16689
	Force	0.19291	0.19935	0.23835
GeoRecon (Ours)	Energy	**0.04646**	**0.07252**	**0.12662**
	Force	**0.15367**	**0.17117**	**0.22662**

GeoRecon achieves lower energy and force errors than its backbone across all temperatures, demonstrating improved stability under distribution shift. These results confirm that reconstruction-based pretraining enhances generalization from small equilibrium molecules to complex, flexible drug-like systems, highlighting the broader applicability of GeoRecon beyond standard benchmarks.

I. Additional Evaluations on Recent Larger Datasets

To evaluate whether the gains extend beyond QM9 and MD17, we include additional matched-setting finetuning results on NablaDFT-Tiny, derived from the ∇^2DFT benchmark (Khrabrov et al., 2024), and PubChemQCR-S (Fu et al., 2025). Coord and GeoRecon use identical finetuning settings within each dataset and split.

Table 16. Energy and force MAE (↓) on larger and more recent benchmarks. GeoRecon consistently improves over Coord under matched finetuning settings.

Dataset / Setting	Model	Energy	Force
NablaDFT-Tiny	SchNet	5.30	56.55
	PaiNN	5.13	46.34
	Coord	5.60	45.12
	GeoRecon	**5.23**	**42.25**
PubChemQCR-S, ST (tiny)	SchNet	1.17	0.44
	PaiNN	0.82	0.37
	Coord	1.90	1.181
	GeoRecon	**1.87**	**1.106**
PubChemQCR-S, SF (tiny)	SchNet	1.19	0.45
	PaiNN	0.86	0.38
	Coord	1.25	0.881
	GeoRecon	**1.20**	**0.880**
PubChemQCR-S, CF (tiny)	SchNet	0.56	0.32
	PaiNN	0.43	0.23
	Coord	0.78	0.261
	GeoRecon	**0.75**	**0.248**

These results are not intended to replace a full benchmark sweep, but they address whether the improvement is confined to small, saturated datasets. Under the same finetuning protocol, GeoRecon improves over Coord on both energy and force across all settings in Table 16.

J. Transferability of GeoRecon

To further examine the transferability of GeoRecon, we conduct a reconstruction-pretraining experiment based on Uni-Mol. We compare the reconstructed Uni-Mol model (UniMol_Rec) against a faithfully reproduced baseline (Uni-Mol (Reproduced)) across five widely used benchmarks: **FreeSolv** (Mobley & Guthrie, 2014) provides hydration free energies for small molecules, reflecting solvation effects in aqueous environments. **Lipo** (Wu et al., 2018) contains experimentally measured octanol–water partition coefficients ($\log P$), a key indicator of molecular permeability and bioavailability in drug discovery. **QM7** (Blum & Reymond, 2009; Rupp et al., 2012) includes atomization energies of 7,165 molecules with up to 23 atoms, computed using density functional theory (DFT), and serves as an early benchmark for energy prediction. **QM8** (Ruddigkeit et al., 2012; Ramakrishnan et al., 2015) reports electronic spectra and excited-state properties of ∼22k molecules, testing the capacity of models to capture quantum phenomena beyond ground-state energies. For **QM9** (Ramakrishnan et al., 2014), we follow the UniMol setting and report results on the HOMO–LUMO gap prediction task.

Table 17. MAE (↓) comparison between UniMol (Reproduced) and UniMol_Rec.

Dataset	UniMol (Reproduced)	UniMol_Rec
FreeSolv	1.6869	**1.6341**
Lipo	**0.6121**	0.6230
QM7	47.1637	**44.7910**
QM8	0.0157	**0.0155**
QM9	0.0046	**0.0046**

Overall, UniMol_Rec yields noticeable gains on FreeSolv and QM7, while maintaining competitive or slightly better performance on QM8 and QM9. These findings suggest that incorporating reconstruction pretraining improves robustness across tasks without introducing extra complexity, underscoring the adaptability of GeoRecon as a transferable paradigm within existing molecular learning frameworks.

K. Effect of Pretraining Dataset Scale on Performance

The size of the pretraining dataset is a critical factor influencing the effectiveness of self-supervised molecular representation learning. To investigate the scalability of GeoRecon, we conducted controlled experiments by pretraining from scratch on two subsets of the PCQM4Mv2 dataset containing 10k and 100k molecules, respectively. After pretraining, the models were finetuned on the QM9 enthalpy prediction

task.

We systematically varied the encoder depth ($L = 8, 10, 12$) and the reconstruction noise scaling factor $\lambda \in \{1.00, 1.05\}$, and report the mean absolute error (MAE) in Table 18.

Table 18. MAE (↓) on QM9 enthalpy(H) after pretraining GeoRecon on PCQM4Mv2 subsets of different sizes.

λ / Encoder	10k Subset			100k Subset		
	$L = 8$	$L = 10$	$L = 12$	$L = 8$	$L = 10$	$L = 12$
1.00	8.438	7.074	7.572	6.109	**5.623**	6.220
1.05	–	7.276	7.360	6.608	5.936	6.089

The results reveal two main observations. First, dataset size substantially impacts pretraining quality: models pretrained on the 100k subset achieve consistently lower errors than those trained on the 10k subset, indicating that additional molecular diversity enhances representation learning. Second, GeoRecon remains effective even in low-resource regimes; when pretrained with only 10k samples, the performance degradation is moderate, and the model still outperforms several full-data baselines. Notably, the 10-layer encoder with $\lambda = 1.00$ on the 100k subset achieves the best finetuning performance, surpassing MolSpectra and Coord models pretrained on the full PCQM4Mv2 dataset.

These findings show GeoRecon is robust to reduced dataset scale while benefiting from additional data, making it attractive for scenarios where only limited pretraining resources are available.

L. Failure Case and Scale Analysis

We analyze failure cases on QM9 using a reproduced Coord baseline under the same evaluation protocol.

Table 19. QM9 error by molecular size. Difference is GeoRecon MAE minus Coord MAE, so negative values indicate GeoRecon improves over Coord.

Size group	Num.	GeoRecon MAE	Coord MAE	Difference
3–9 atoms (small)	23	0.2095	0.2822	-0.0727
10–15 atoms	2094	0.0200	0.0256	-0.0056
16–20 atoms	6499	0.0216	0.0263	-0.0047
21+ atoms (large)	2215	0.0249	0.0291	-0.0042

GeoRecon improves over Coord across all size regimes. Errors are higher for very small and larger molecules, but the improvement is not restricted to one size group; the largest absolute improvement appears for the small-molecule group.

Compared with Coord, GeoRecon shows stronger dependence on geometric factors but equal or lower dependence on topology, size, and flexibility. This suggests that remaining errors are more closely tied to geometric heterogeneity than to molecule size alone.

Table 20. Pearson correlations between prediction error magnitude and molecular factors. Lower average absolute correlation indicates weaker dependence on that factor group.

Factor group	GeoRecon Avg. $\lvert r\rvert$	Coord Avg. $\lvert r\rvert$
Geometry	0.076	0.059
Topology	0.055	0.070
Size	0.074	0.074
Flexibility	0.047	0.049

The largest outlier in this analysis is H_2O, whose ground-truth value is 6.002. GeoRecon predicts 9.563, while Coord predicts 11.762. Removing this single case gives GeoRecon an MAE of 0.0220 versus 0.0267 for Coord. Remaining GeoRecon failure cases are concentrated in fluorinated small molecules and alkyne chains, such as `[H]C#CC#CC(F)(F)F`, while GeoRecon improves over Coord on several molecules with N/O heteroatoms and amide-like structures, such as `[H]N([H])C1C2=C(NN=N2)C([H])([H])N1[H]`. These cases indicate concrete directions for future improvements in geometry-aware pooling and local-global interaction modeling.

M. Baselines

Training from scratch: SchNet (Schütt et al., 2018) employs continuous-filter convolutions for quantum interactions. EGNN (Satorras et al., 2021) introduces equivariant message passing under Euclidean transformations. DimeNet (Gasteiger et al., 2020b) and DimeNet++ (Gasteiger et al., 2020a) leverage angular information, with the latter improving efficiency and accuracy. PaiNN (Schütt et al., 2021) extends equivariant message passing to tensorial targets. SphereNet (Liu et al., 2021) incorporates spherical coordinates, while TorchMD-Net (Thölke & De Fabritiis, 2022) applies equivariant transformers to molecular dynamics.

Pretraining then finetuning: Transformer-M (Luo et al., 2023) encodes both 2D and 3D molecular data. SE(3)-DDM (Liu et al., 2023b) adopts SE(3)-invariant denoising distance matching. 3D-EMGP (Jiao et al., 2023) leverages energy-based equivariant objectives. Coord (Zaidi et al., 2023) applies coordinate denoising for geometry-aware learning. Frad (Feng et al., 2023) combines dihedral and coordinate perturbations through fractional denoising. SliDe (Ni et al., 2024) uses physics-informed sliced denoising over internal coordinates. DenoiseVAE (Liu et al., 2025b) learns molecule-adaptive noise distributions, while AniDS (Liu et al., 2025a) learns anisotropic atom-specific covariance structures. AniDS w/ GeoRecon adds graph-conditioned reconstruction to AniDS under the same

matched-evaluation principle.

Graph-level and multimodal pretraining baselines: MoleculeSDE (Liu et al., 2023a), included in the QM9 main table, introduces group-symmetric SDE pretraining to align molecular modalities. MolSpectra (Wang et al., 2025), used as a full-data reference in the dataset-scale analysis, uses multimodal energy spectra as an additional quantum-informed signal. These methods are included to make the graph-level baseline coverage explicit and to position GeoRecon relative to prior molecule-level or multimodal supervision.

N. Training Information

Devices. Experiments on MD17 and QM9 were conducted using NVIDIA RTX 4090 GPUs. Pretraining was performed on 4 × RTX 4090 GPUs (24 GB each) for 400k steps, while downstream finetuning for each task was carried out on a single RTX 4090 GPU. For MD22, finetuning experiments were conducted on a single NVIDIA A100 GPU.

Hyperparameter Settings. Our hyperparameter settings are shown in Table 21. They closely follow Coord (Zaidi et al., 2023), with one modification: we reduce the QM9 pretraining batch size from 70 to 50 due to GPU memory constraints.

Table 21. Training Configuration

	Finetuning		**Pretraining**	
	QM9	MD17	QM9	MD17
batch_size	128	4	50	70
cutoff_lower	0	0	0	0
cutoff_upper	5	5	5	5
ema_alpha_dy	1	1	1	1
ema_alpha_y	1	0.05	1	1
embedding_dimension	256	128	256	128
energy_weight	1	0.2	0	0
force_weight	1	0.8	1	1
inference_batch_size	128	64	70	70
lr	0.0004	0.0005	0.0004	0.0004
lr_schedule	cosine_warmup	cosine_warmup	cosine	cosine
lr_min	1e-7	1e-7	1e-7	1e-7
lr_patience	15	30	15	15
lr_warmup_steps	10000	1000	10000	10000
lr_cosine_length	100000	20000	400000	400000
max_num_neighbors	32	32	32	32
max_z	100	100	100	100
num_heads	8	8	8	8
num_layers	8	6	8	6
num_nodes	1	1	1	1
num_rbf	64	32	64	32
num_workers	6	6	6	6
precision	32	32	32	32
save_interval	10	10	1	1
test_interval	10	100	1	1
position_noise_scale	0	0.005	0.04	0.04
denoising_weight	0.1	0.1	1	1

Computational Cost. We evaluated three models–Coord, GeoRecon, and Frad–under the same experimental setup using the TorchMD-Net backbone on a single NVIDIA RTX 4090 (24GB) GPU with a batch size of 25. The average per-step training times were 0.1057 s for Coord, 0.1867 s for GeoRecon, and 0.2200 s for Frad. Although GeoRecon incurs a higher training cost than Coord, it remains more efficient than Frad. The additional overhead primarily arises from encoding three molecular variants (clean, noised, and large-noise) at each training step, whereas Coord processes only the noised conformation. Importantly, this extra cost is incurred only during the one-time pretraining phase. Since the pretrained checkpoint can be reused across diverse downstream tasks, the practical impact of this overhead is limited, while the benefits in downstream performance are more consequential.

O. Pooling Strategies for Graph-Level Conditioning

GeoRecon conditions the reconstruction head on a graph-level vector aggregated from node representations. Mean pooling is used by default due to its simplicity and stability under coordinate perturbations, but other permutation-invariant aggregators are compatible. In preliminary experiments, we also considered sum pooling and attention-based pooling; these alternatives did not yield consistent improvements over mean pooling under the same training budget, while introducing additional parameters and sensitivity to optimization. We therefore adopt mean pooling throughout.

The released code will be made publicly available under the MIT License, while third-party assets retain their original licenses.