
Force Alignment Does Not Predict Transfer: Contrasting Denoising and Force Regression

Max Rausch-Dupont & Dietrich Klakow
Spoken Language Systems
Saarland University
Saarbrücken, SL 66123, GER
{mrdupont,klakow}@lsv.uni-saarland.de

Abstract

Recent approaches to pre-training atomistic neural networks refine the denoising objective to improve the implicitly learned force-field, assuming a direct link between force-field accuracy and downstream gains on property prediction. We test this premise with a matched comparison of denoising against supervised force regression, relating output and representation-level force alignment to fine-tuning performance. While denoising yields poorer and degrading force alignment, it still achieves decreasing downstream error and maintains an advantage over force regression. These results question whether explanations of downstream gains should be based on the accuracy of the learned force field.

1 Introduction

Molecular property prediction is limited by label cost and dataset size, which motivates self-supervised pre-training, an approach well established in vision and language [2, 4, 6, 16]. The forces acting on atoms and numerous properties derive from the same potential energy surface, so a model trained on forces may transfer to property prediction. For structures drawn from the Boltzmann distribution those forces are the *score* of that distribution, and such a score can be estimated without labels by denoising [21]. Zaidi et al. [23] therefore pre-train by *coordinate denoising*, adding isotropic Gaussian noise to an equilibrium conformation and training a network to predict the perturbation.

Follow-up work [5, 11, 12, 14] proposes perturbations more faithful to the anisotropic potential energy surfaces encountered in practice, and the resulting downstream gains are presented as evidence that *force-field learning is what makes denoising work*. This hypothesis remains empirically unverified, and alternative explanations exist. For instance, Perez & Gomez-Bombarelli [15] conjecture that denoising primarily fosters smoother, more distinguishable atomic embeddings. We investigate this premise in a controlled setting by contrasting a denoising model with a force regression model trained on identical data. Force regression learns the force-field by construction and therefore helps isolate what downstream gains can be attributed to force accuracy. We compare the forces they produce, the force information they retain internally, and how well either predicts finetuning performance.

Contributions According to Vincent [21] denoising approximates a force-field, but: (1) its output aligns far less with reference forces than force regression does; (2) alignment degrades both with depth in the network and over pre-training; (3) downstream error decreases regardless, while alignment is flat or falling; and (4) despite its less accurate forces, denoising outperforms force regression at every checkpoint we finetuned. Force alignment therefore neither predicts transfer nor provides a criterion for when to stop pre-training, and the benefit of pre-training should not be attributed to the accuracy of the learned forces alone.

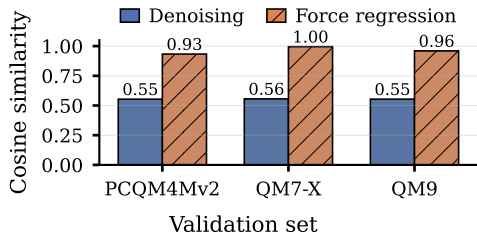


Figure 1: Average cosine similarity between the model output and the reference forces of each dataset, computed at its own level of theory (appendix B). Denoising yields a much lower similarity than direct force prediction.

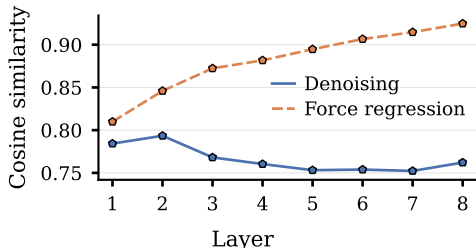


Figure 2: Linear probes on the layer representations after pre-training. Average cosine similarity between probe prediction and target forces decreases with depth for the denoiser. The force-supervised model shows the opposite behaviour.

2 Preliminaries & setup

Denoising In coordinate denoising [23], a conformation $\mathbf{x} \in \mathbb{R}^{n \times 3}$ (a local minimum of the potential energy surface U) is perturbed with isotropic Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. A denoising network r is then trained to predict the noise ϵ from the corrupted structure $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$ by minimizing the squared error $\|\epsilon - r(\tilde{\mathbf{x}}; \theta)\|_2^2$. Following previous work [5, 14, 23], we instantiate r with the equivariant Transformer [19] and keep $\sigma = 0.04$ throughout. We provide results for different noise scales in appendix C.5.

Connecting denoising to force regression Structures at thermal equilibrium are Boltzmann-distributed, $p(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$, so the score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ is the force $\mathbf{F}(\mathbf{x}) = -\nabla_{\mathbf{x}} U(\mathbf{x})$. Denoising is equivalent to *score matching* [9] against $q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2 \mathbf{I})$ [21]. The score of q_{σ} has the closed form $\nabla_{\tilde{\mathbf{x}}} \log q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x}) = -\epsilon/\sigma^2$. The trained network therefore implies the score estimate $-r(\tilde{\mathbf{x}}; \theta)/\sigma^2$. Crucially, q_{σ} is the distribution of perturbations around a single conformation, not the Boltzmann distribution: its score $(\mathbf{x} - \tilde{\mathbf{x}})/\sigma^2$ is the force of an isotropic harmonic well of curvature $\mathbf{I}/\sigma^2$ centred at $\mathbf{x}$, which agrees with the force of the true potential U only if U is locally of that same curvature. In that case denoising coincides with supervised force regression (appendix D).

Setup PCQM4Mv2 [13], the standard pre-training corpus in this line of work [5, 11, 12, 14, 23], provides only equilibrium structures and no reference forces and therefore does not support a matched comparison against force regression. We instead pre-train both objectives on QM7-X [7], which supplies equilibrium structures, off-equilibrium structures close to them, and the forces acting at those off-equilibrium positions. QM7-X contains several perturbed structures per conformation, of which we use one at random, so both objectives see the same molecules, though not necessarily the same structures. For denoising we also pre-train on PCQM4Mv2 to connect our results to the established setting [5, 12, 10, 14, 23]. To evaluate both models beyond the chemical space of QM7-X, we sample one thousand structures each from PCQM4Mv2 and QM9 [17], perturb them with isotropic Gaussian noise at the pre-training noise scale, and compute the acting forces at the level of theory of their reference geometries (appendix B). Unlike our isotropically perturbed reference forces, the off-equilibrium structures in QM7-X come from normal-mode sampling around an equilibrium structure (full harmonic approximation). Downstream, we finetune on the U_0 target of QM9 over three random train/validation splits with a shared held-out test set. Full details are given in appendix A.

3 Results

3.1 Do the two objectives produce the same forces?

In theory, both objectives target the same force-field, so we first compare their *force alignment*, the average cosine similarity between predicted and reference forces. We train one model to denoise perturbed equilibrium structures in QM7-X, and another to predict the forces of the off-equilibrium structures already in the dataset. Evaluating both on the QM7-X off-equilibrium structures favours

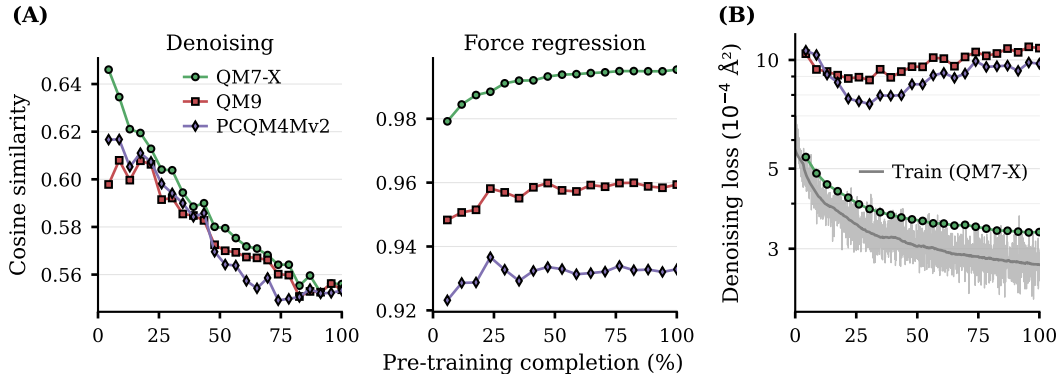


Figure 3: (A) Average cosine similarity of model output to reference forces during pre-training. (B) Denoising MSE loss on the QM7-X training set (solid, rolling average) and on the QM7-X, QM9 and PCQM4Mv2 validation sets (marked lines). Loss on QM9 and PCQM4Mv2 eventually diverges.

force regression, which was trained on normal-mode samples, over the denoiser, which only ever saw isotropically perturbed ones. We therefore also evaluate both models on unseen isotropically perturbed structures from PCQM4Mv2 and QM9, which neither model has been trained on.

Results Across datasets, denoising reaches a substantially lower force alignment than supervised force regression (fig. 1). To see whether denoising actively discards this information, we track force alignment during pre-training. As fig. 3 (A) shows, denoising harms the alignment with the reference forces, whereas force regression improves continuously on all datasets.

3.2 Do the two objectives retain the same information?

Since the final layers of a network are known to specialize to the objective they were trained on [22, 18], the lower force accuracy of the denoiser need not imply that the information is absent from its representations. To test this, we fit linear probes [1] on the vectorial representations of both models, five times per layer on random train/test splits drawn from our QM7-X validation set. Figure 2 shows the average cosine similarity of the probe’s output to the reference forces.

Results Contrary to our expectation, a gap remains between the two models, with the information more accessible in the model trained by force regression. Probes with more predictive power behave the same way, and probe accuracy also decreases during pre-training (appendix C.1). The two objectives target the same force-field, yet the models differ both in output and in what is accessible inside them. We next test whether this difference carries over to downstream performance.

3.3 Does force alignment predict transfer?

If the benefit of pre-training rests on the force-field the objective learns, a model whose output aligns better with the reference forces should transfer better. We therefore checkpoint periodically during pre-training and finetune at 5, 10, 15, 35, 65, 85, and 100% completion on the U_0 target of QM9 [17]. For comparison we also train a model from scratch.

Results The from-scratch model reaches a test mean absolute error (MAE) of 6.05 meV. Table 1 reports the improvement of the finetuned models over this baseline in percent. We find that either pre-training objective leads to a reduction in the test MAE, with denoising yielding the greater improvement at every checkpoint. This cannot be solely explained by the number of distinct structures seen: drawing a new perturbed structure in every epoch leaves the downstream error of force regression mostly unchanged (appendix C.4). Similar results of self-supervised representations and pre-training objectives outperforming a supervised alternative have been observed in other domains [2, 3, 6, 20]. The force alignment of the denoising model is highest at the beginning of pre-training and declines from there on (fig. 3, A), yet its earliest checkpoint already beats the from-scratch baseline and later, less aligned ones improve further. Similarly, the force alignment of

Table 1: Improvement in QM9 test MAE over the 6.05 meV from scratch baseline in percent, against pre-training completion. Mean $\pm$ SD over three finetuning seeds, the baseline is a single run. DN = denoising, FR = force regression. Absolute values are shown in table 3.

Pre-training	Pre-training completion (%)						
	5	10	15	35	65	85	100
QM7-X, FR	2.6 $\pm$ 1.2	4.9 $\pm$ 2.1	4.5 $\pm$ 3.7	5.7 $\pm$ 1.7	8.5 $\pm$ 3.0	8.1 $\pm$ 0.8	12.6 $\pm$ 2.4
QM7-X, DN	9.9 $\pm$ 2.1	6.9 $\pm$ 3.7	7.4 $\pm$ 3.0	7.4 $\pm$ 1.7	12.3 $\pm$ 4.2	13.7 $\pm$ 3.4	18.8 $\pm$ 0.9
PCQM4Mv2, DN	7.1 $\pm$ 0.7	6.9 $\pm$ 1.3	12.6 $\pm$ 3.1	15.9 $\pm$ 2.4	20.3 $\pm$ 3.5	23.3 $\pm$ 1.3	23.2 $\pm$ 0.9

the force-regression model plateaus after roughly 35% of pre-training, yet the finetuned models keep improving. Force alignment thus seems to have no direct impact downstream.

While the force alignment of the model output generalizes from the training dataset (QM7-X) to the other datasets (fig. 3, A), the same cannot be said for denoising. Measuring how well the model denoises perturbed structures of PCQM4Mv2 and the downstream dataset, we find that it first improves everywhere and then diverges (fig. 3, B). This is a classic sign of overfitting, nonetheless the downstream error continues to decrease.

This divergence does not appear on every pre-training dataset. The larger PCQM4Mv2 dataset reproduces the decrease in force alignment, but the denoising loss on QM9 follows the training loss throughout (fig. 9), while the downstream error improves in both cases.

Considering these results, it seems that neither the denoising loss, nor the similarity to reference forces can be taken as robust measures of how well the model will perform downstream. Downstream error nonetheless improves with pre-training progress for both objectives, so what transfers to finetuning is something other than what the model’s the pre-training objective measures.

4 Conclusions

We set out to test whether the downstream benefit of denoising pre-training can be attributed to the force-field the objective implicitly learns. Comparing denoising against direct force regression under matched data, we find that the two objectives, equivalent in theory, behave differently in practice. The denoiser’s output aligns considerably less with reference forces, and this alignment degrades over pre-training instead of improving. The same holds for the accessibility of force information in the intermediate representations, which decreases with network depth rather than increasing. Neither effect carries over to finetuning. Downstream error continues to decrease with pre-training progress for both objectives, and it does so in regimes where the force alignment is flat or falling and where the denoising loss on the downstream distribution has already started to diverge.

Within the scope of our experiments, these results are difficult to reconcile with the view that the accuracy of the learned force-field is what makes denoising pre-training work. Downstream error falls while alignment degrades, and denoising, despite its less accurate forces, outperforms force regression downstream. Furthermore, neither force alignment nor denoising validation loss tracks downstream success in our settings, so neither is a safe early-stopping criterion.

What pre-training changes therefore remains open. Since it is evidently not the fidelity of the learned force-field, a natural next step is to characterize which property of the pre-trained network survives finetuning, and whether it is the same for both objectives. Such a property appears to be present already after 5% of pre-training, well before either pre-training loss has converged.

5 Limitations

We consider only the simplest form of denoising, isotropic Gaussian perturbations at a single noise scale, with a single backbone. Anisotropic methods like SliDe [14] improve downstream results and are expected to better approximate reference forces, so their gains may correlate with force approximation accuracy, which makes the anisotropic setting a compelling area for further study. We also measure downstream performance on a single target, U_0 of QM9, on a chemical space close to

that of QM7-X. U_0 is dominated by the value of the potential energy surface at a minimum, where the forces vanish, so it depends on the learned force-field only indirectly. Other QM9 targets, such as the zero-point vibrational energy and the heat capacity, depend directly on the curvature of the surface at the minimum, so force alignment may matter more for them. Our results show that force alignment does not predict transfer to U_0 , we leave other targets to future work.

Acknowledgments and Disclosure of Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914. We thank Aravind Krishnan and Andreas Rouvalis for helpful feedback.

References

- [1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=HJ4-rAVt1>.
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, October 2021. doi: 10.1109/ICCV48922.2021.00951.
- [3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9620–9629. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00950. URL <https://doi.org/10.1109/ICCV48922.2021.00950>.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- [5] Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional Denoising for 3D Molecular Pre-training. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 9938–9961. PMLR, July 2023.
- [6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL <https://doi.org/10.1109/CVPR52688.2022.01553>.
- [7] Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio Jr., and Alexandre Tkatchenko. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data*, 8(1):43, February 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00812-2.
- [8] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/db8e1af0cb3aca1ae2d0018624204529-Abstract-round2.html>.
- [9] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. URL <https://jmlr.org/papers/v6/hyvarinen05a.html>.

- [10] Yi-Lun Liao, Tess E. Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *Trans. Mach. Learn. Res.*, 2024, 2024. URL <https://openreview.net/forum?id=whGzYUbIWA>.
- [11] Xixian Liu, Rui Jiao, Zhiyuan Liu, Yurou Liu, Yang Liu, Ziheng Lu, Wenbing Huang, Yang Zhang, and Yixin Cao. Learning 3D Anisotropic Noise Distributions Improves Molecular Force Fields. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.
- [12] Yurou Liu, Jiahao Chen, Rui Jiao, Jiangmeng Li, Wenbing Huang, and Bing Su. Denoise-VAE: Learning Molecule-Adaptive Noise Distributions for Denoising-based 3D Molecular Pre-training. In *The Thirteenth International Conference on Learning Representations*, October 2024.
- [13] Maho Nakata and Tomomi Shimazaki. PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, June 2017. ISSN 1549-9596. doi: 10.1021/acs.jcim.7b00083.
- [14] Yuyan Ni, Shikun Feng, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Sliced Denoising: A Physics-Informed Molecular Pre-Training Method. In *The Twelfth International Conference on Learning Representations*, October 2023.
- [15] Tynan Perez and Rafael Gomez-Bombarelli. Self-Conditioned Denoising for Atomistic Representation Learning, March 2026.
- [16] Alec Radford and Karthik Narasimhan. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI, Inc, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [17] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22.
- [18] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1452. URL <https://doi.org/10.18653/v1/p19-1452>.
- [19] Philipp Thölke and Gianni De Fabritiis. Equivariant Transformers for Neural Network based Molecular Potentials. In *International Conference on Learning Representations*, October 2021.
- [20] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Holger Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, July 2022.
- [21] Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00142. URL https://doi.org/10.1162/NECO_a_00142.
- [22] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3320–3328, 2014. URL <https://proceedings.neurips.cc/paper/2014/hash/375c71349b295f5be2dcdca9206f20a06-Abstract.html>.
- [23] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter W. Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=tYIMtogyee>.

A Training details

A.1 Architecture

All experiments use the TorchMD-Net equivariant transformer (ET) of Thölke & Fabritiis [19] as backbone, in the same configuration for pre-training and fine-tuning, so that a pre-trained encoder can be transferred without any change. The encoder maps a molecule (z, x) to per-atom scalar features $h \in \mathbb{R}^{n \times d}$ and vector features $V \in \mathbb{R}^{n \times 3 \times d}$, with $d = 256$ hidden channels and 8 layers of equivariant multi-head attention (8 heads). Interatomic distances are expanded in 64 exponential-normal radial basis functions (not trainable) over the interval $[0, 5]$ Å, with a maximum of 64 neighbours per atom, and both the attention weights and the vector features are cut off smoothly at 5 Å. Activations are SiLU throughout, including inside the attention, and a neighbour embedding is applied to the initial atom-type embedding ($Z_{\max} = 100$). This gives 6.52M parameters in the encoder, and 6.87M including the output head.

Two output heads are used on top of the encoder, both built from gated equivariant blocks that mix h and V . The *vector* head (EquivariantVectorOutput) returns one 3-vector per atom, and serves both as the denoising head, predicting the position correction, and as the direct force head. The *scalar* head (EquivariantScalar, two gated equivariant blocks of widths $256 \rightarrow 128 \rightarrow 1$) returns a per-atom scalar that is summed over the molecule; it is used for the downstream U_0 regression, on top of an Atomref prior that supplies the per-element reference energies.

A.2 Pre-training

Pre-training is run with Adam at a constant learning rate of 10^{-4} , batch size 32, in single precision on one GPU, for at most 250 epochs with early stopping (patience 10 validation checks) on the objective’s own validation loss. Checkpoints are written every 5000 optimizer steps. We report results for fine-tuning at 5, 10, 15, 35, 65, 85 and 100% of the completed pre-training run.

For the denoising objective, positions are corrupted with isotropic Gaussian noise of standard deviation $\sigma = 0.04$ Å (the value reported by Ni et al. [14] to work best), drawn independently per atom and per epoch. The model is trained with the reconstruction parameterization $r_x(\tilde{x}) = \tilde{x} + \text{ET}(\tilde{x})$ against the clean positions, which is equivalent to noise prediction by eq. (1). To compute the force-alignment we use the pseudo-force $\text{ET}(\tilde{x})/\sigma^2$ implied by eq. (2), evaluated at the same noise scale it was trained at.

Force regression differs only in its loss: the same vector head is regressed non-conservatively onto the reference forces of the pre-training corpus. On QM7-X denoising is trained on the equilibrium structures and the force regression uses one displaced structure per equilibrium, so the objectives see the same molecules and the same number of training examples.

A.3 Fine-tuning

All downstream results are QM9 U_0 regression. The pre-trained encoder weights are transferred, the pre-training head is discarded, and a freshly initialized EquivariantScalar head with an Atomref prior is trained jointly with the encoder. We train with AdamW (weight decay 0) at a peak learning rate of 4×10^{-4} with 10^4 linear warm-up steps, followed by ReduceLROnPlateau (factor 0.8, patience 15 epochs, floor 10^{-7}), batch size 128, for at most 3000 epochs with early stopping after 150 epochs without improvement. Training loss is mean squared error on U_0 , while checkpoint selection and early stopping monitor the validation mean absolute error.

QM9 is split 110,000/10,000/10,831 into train, validation and test. Train, validation set and weight initialization vary with the seeds (0, 1, 2), the test partition is held fixed across all seeds and runs, so test MAE remains comparable.

B Datasets

We use three corpora: PCQM4Mv2 and QM7-X for pre-training, and QM9 for fine-tuning.

Unless stated otherwise a dataset is split 80/10/10 into train, validation and test from a fixed seed. For PCQM4Mv2 the draw is a uniform random permutation over molecules. For QM7-X it is *grouped* by the molecule identifier idmol, so that every structure belonging to one molecule falls in the same

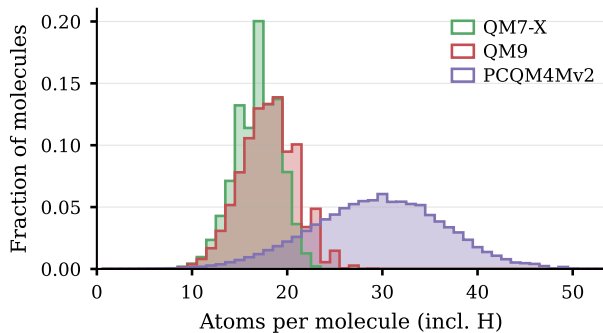


Figure 4: Distribution of the number of atoms per molecule, hydrogens included.

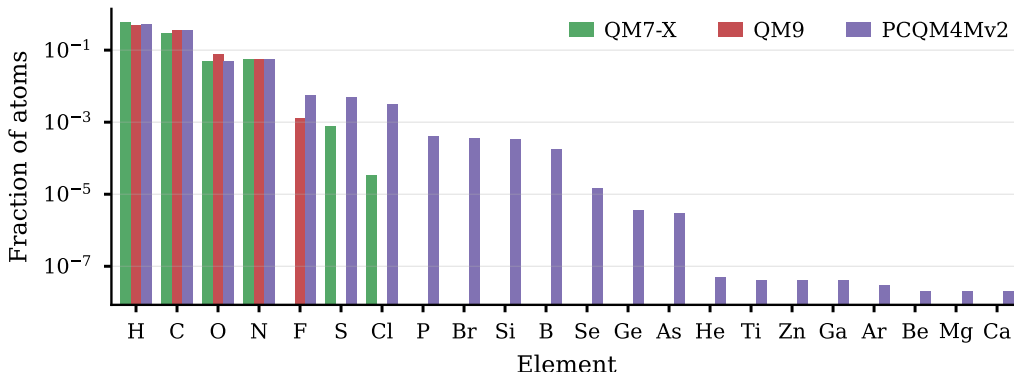


Figure 5: Fraction of all atoms in each corpus that belong to a given element, on a logarithmic scale.

partition and no model is validated on a displaced copy of a structure it was trained on. QM9 uses a stratified split for the test set. Resulting partition sizes are listed in table 2. Figures 4 and 5 compare the three corpora in molecule size and element composition.

Table 2: Datasets after our preprocessing, with the level of theory the stored geometries were produced at and our train/validation/test partition sizes. For QM7-X the counts refer to the equilibrium subset actually used for pre-training (appendix B.2).

Dataset	Geometries	Level of theory	Partition sizes		
			Train	Val	Test
PCQM4Mv2	3,378,573	B3LYP/6-31G*	2,702,859	337,857	337,857
QM7-X (eq.)	39,637	DFTB3+MBD	31,709	3,965	3,963
QM9	130,831	B3LYP/6-31G(2df,p)	110,000	10,000	10,831

B.1 PCQM4Mv2

PCQM4Mv2 is the quantum-chemistry subset of PubChemQC [13], redistributed by OGB-LSC as a benchmark [8]. It is the standard pre-training corpus in the denoising literature [5, 11, 12, 14, 23]. We use the training split of the v2 release, the only one for which the 3D structures are public: one B3LYP/6-31G* optimised equilibrium geometry per molecule. There are no forces and no off-equilibrium structures, which is why the matched comparison against force regression cannot be run on this corpus. We discard 33 entries RDKit fails to parse, leaving 3,378,573 molecules. Molecules are considerably larger than those of QM7-X and QM9, which makes this a corpus that is out of distribution with respect to the downstream task.

B.2 QM7-X

QM7-X [7] covers the molecules of GDB-13 with up to seven non-hydrogen atoms drawn from C, N, O, S and Cl. Structures are optimised at the DFTB3+MBD level, and off-equilibrium structures are obtained by normal-mode sampling around each optimised geometry. Forces are recomputed at the PBE0+MBD level for every structure, equilibrium and off-equilibrium alike, which is what makes QM7-X a corpus that supports supervised force regression at the same structures a denoiser would be trained on. PBE0+MBD forces are converted to kcal/mol/Å.

Excluded structures. We exclude the duplicate structures flagged in the release’s `DupMols.dat`, together with the off-equilibrium structures displaced from them. This leaves 39,637 optimised structures and 4,003,337 structures in total.

Structures used for pre-training. For denoising we use all 39,637 optimised structures. For force regression we sample, for each of them, one of its 100 off-equilibrium structures uniformly at random, so that both objectives are trained on 39,637 structures of the same molecules. For the experiments in appendix C.4 we randomly pick one of the 100 structures for training.

B.3 QM9

QM9 [17] holds 133,885 molecules with up to nine non-hydrogen atoms drawn from C, N, O and F, enumerated from GDB-17, with geometries optimised at the B3LYP/6-31G(2df,p) level and twelve regression targets per molecule. We remove the 3,054 molecules listed in the release’s own `uncharacterized.txt`, whose geometry optimisation failed a consistency check, leaving 130,831 molecules. All downstream results in this paper are on the atomisation energy U_0 .

Splitting. We keep the 110,000/10,000 train/validation sizes that are conventional for this benchmark, so that our from-scratch baseline and our fine-tuned models are comparable to numbers reported elsewhere, and let the test partition absorb the remaining 10,831 molecules. The test partition is drawn once from a fixed seed, stratified over U_0 into ten quantile bins so that it covers the target range, and is then held out identically across every run in this paper. For each fine-tuning seed the rest is split uniformly at random into train and validation.

B.4 Force computation

Neither PCQM4Mv2 nor QM9 ships off-equilibrium structures, so we compute reference forces at *perturbed* geometries that are required for force-alignment metrics.

For each dataset we draw 1000 molecules uniformly without replacement from the *validation* split. Each sampled equilibrium geometry is perturbed once with isotropic Gaussian noise of standard deviation $\sigma = 0.04$ Å, and the reference force is the negative energy gradient evaluated at that perturbed geometry. Sampling and noise are drawn from a fixed seed, so all models are scored against exactly the same structures and the same targets.

Each molecule is treated at the level of theory its own dataset’s geometries were produced at: B3LYP/6-31G(2df,p) for QM9 and B3LYP/6-31G* for PCQM4Mv2. Computations are performed with PySCF 2.13.0. All molecules are treated as neutral and the spin multiplicity is taken from the electron-count parity, which selects restricted Kohn–Sham for the closed-shell cases and unrestricted Kohn–Sham for the rest. We converge the self-consistent field (SCF) to 10^{-8} Hartree within at most 200 cycles and apply a level shift of 0.2; everything else is left at PySCF’s defaults. Molecules whose SCF does not converge are discarded. Gradients are converted from Hartree/Bohr to kcal/mol/Å and negated, which puts them in the same units as the QM7-X reference forces, so cosine similarities are directly comparable across corpora.

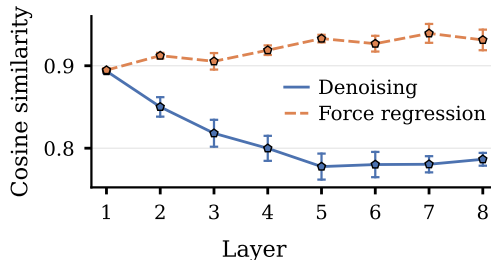


Figure 6: Cosine similarity of an equivariant MLP probe to QM7-X forces across layers.

C Extended results

C.1 Probing

This section describes how the probes of fig. 2 are fitted and reports the probing results referenced in the main text in full.

Probed representation and reference set. Probes are fitted on the encoder’s *vectorial* stream $V \in \mathbb{R}^{n \times 3 \times d}$, against the QM7-X *test* partition of table 2 unless stated otherwise, using the normal-mode displaced structures and their PBE0+MBD forces. Since QM7-X is split grouped by molecule, those structures are disjoint from the pre-training set. All representations are computed with frozen encoder weights, at exactly the geometry the reference force was computed at (no noise is added at probing time).

Probe training. Both probes map one atom’s features $V_i \in \mathbb{R}^{3 \times d}$, with columns v_{ic} , to $f_i = \sum_c w_{ic} v_{ic}$: they contract the channel axis only and never mix x, y and z .

The *linear* probe learns one weight vector $w_{ic} = w_c$, fitted by ridge regression ($\alpha = 1$) without intercept. We fit a single model across all atoms and Cartesian components.

In the *gated* probe an MLP predicts w_{ic} from the channel norms $\|v_{ic}\|$, i.e. TorchMD-Net’s gated equivariant block [19] restricted to the vector stream, with one hidden layer of width $d/4 = 64$ and SiLU. We normalize the input by the root mean square of the channel norms over the training fold, and fit with Adam (learning rate 10^{-2} , no weight decay) on the root mean squared error, in batches of 256 atoms, for at most 500 epochs. Training stops after 25 consecutive epochs that fail to improve on the best epoch so far by 10^{-5} of the first epoch’s loss.

Both probes are cross-validated over 5 folds grouped by molecule, with one probe fitted per layer, family and fold. Targets are divided by the mean force norm of the training fold. Reported values are the mean over the five held-out folds of the per-atom cosine similarity between probe prediction and reference force.

Results We present the equivalent of fig. 2 with a equivariant MLP in fig. 6. We further show the cosine similarity of the probe prediction to the QM7-X reference forces resolved by layer and by pre-training completion, for a linear probe (fig. 7) and for an equivariant MLP (fig. 8).

C.2 PCQM4Mv2 training

For the denoising model pre-trained on PCQM4Mv2, we show the force alignment of the model output and the denoising loss on the training and validation distributions over the course of pre-training, alongside the downstream test error of the corresponding fine-tuned checkpoints in fig. 9.

C.3 U_0 results

Table 3 gives the absolute QM9 U_0 test MAE underlying the relative improvements of table 1, for both pre-training objectives at each checkpoint.

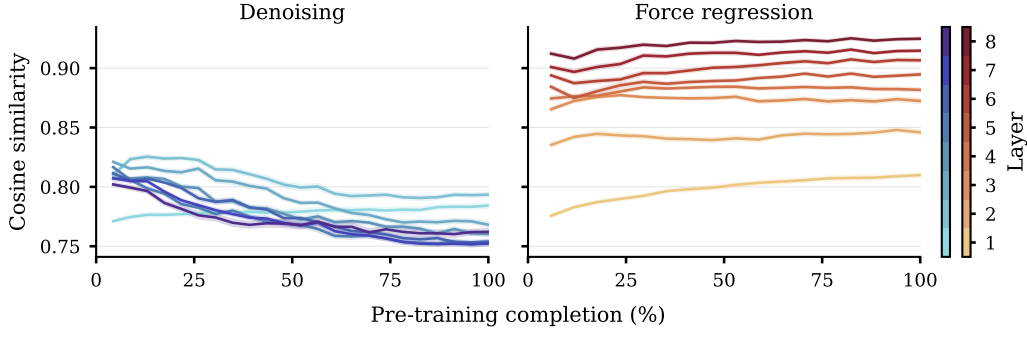


Figure 7: Cosine similarity of a linear probe to QM7-X forces across pre-training and layers.

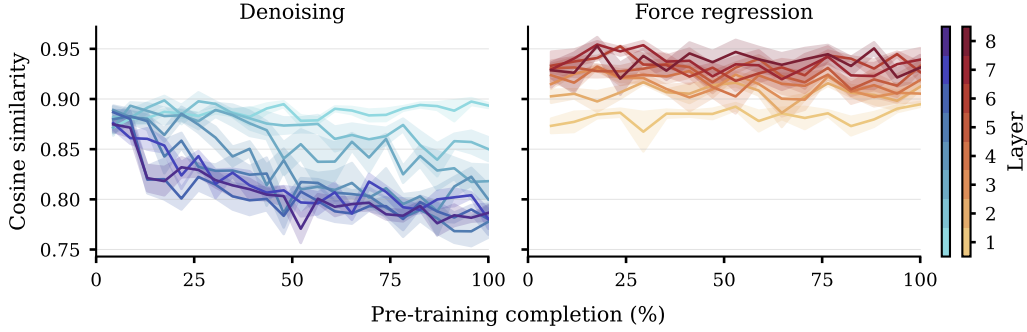


Figure 8: Cosine similarity of an equivariant MLP probe to QM7-X forces across pre-training and layers.

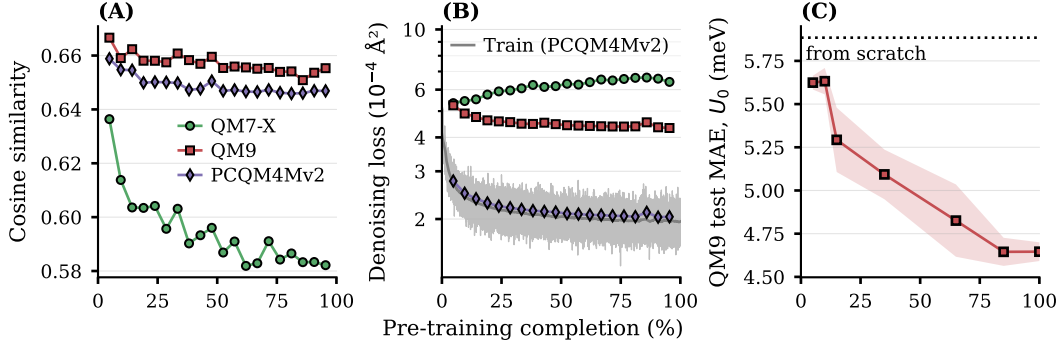


Figure 9: Pre-training by denoising on the PCQM4Mv2 dataset.

Table 3: QM9 U_0 test MAE in meV after fine-tuning, as a function of how far pre-training had run. Mean $\pm$ SD over to 3 fine-tuning seeds. DN = denoising, FR = force regression.

Pre-training	Pre-training completion (%)						
	5	10	15	35	65	85	100
QM7-X, FR	5.90 $\pm$ 0.07	5.75 $\pm$ 0.13	5.78 $\pm$ 0.23	5.71 $\pm$ 0.10	5.54 $\pm$ 0.18	5.56 $\pm$ 0.05	5.29 $\pm$ 0.14
QM7-X, DN	5.45 $\pm$ 0.13	5.64 $\pm$ 0.22	5.60 $\pm$ 0.18	5.60 $\pm$ 0.10	5.31 $\pm$ 0.25	5.22 $\pm$ 0.20	4.92 $\pm$ 0.06
PCQM4Mv2, DN	5.62 $\pm$ 0.04	5.63 $\pm$ 0.08	5.29 $\pm$ 0.19	5.09 $\pm$ 0.14	4.83 $\pm$ 0.21	4.64 $\pm$ 0.08	4.65 $\pm$ 0.05

Table 4: QM9 U_0 test MAE against pre-training completion and the number of distinct structures seen (appendix C.4). Mean $\pm$ SD over three fine-tuning seeds. DN = denoising, FR = force regression. [†]Expected value.

Pre-training run	Completion (%)	Structures (M)	QM9 U_0 test MAE (meV)
PCQM4Mv2, DN	5	2.880	5.6 $\pm$ 0.0
	10	5.600	5.6 $\pm$ 0.1
	15	8.480	5.3 $\pm$ 0.2
	35	19.840	5.1 $\pm$ 0.1
	65	36.800	4.8 $\pm$ 0.2
	85	48.160	4.6 $\pm$ 0.1
	100	56.640	4.6 $\pm$ 0.1
QM7-X, DN	5	0.160	5.5 $\pm$ 0.1
	10	0.320	5.6 $\pm$ 0.2
	15	0.480	5.6 $\pm$ 0.2
	35	1.280	5.6 $\pm$ 0.1
	65	2.400	5.3 $\pm$ 0.3
	85	3.200	5.2 $\pm$ 0.2
	100	3.680	4.9 $\pm$ 0.1
QM7-X, FR	5	0.032	5.9 $\pm$ 0.1
	10	0.032	5.8 $\pm$ 0.1
	15	0.032	5.8 $\pm$ 0.2
	35	0.032	5.7 $\pm$ 0.1
	65	0.032	5.5 $\pm$ 0.2
	85	0.032	5.6 $\pm$ 0.1
	100	0.032	5.3 $\pm$ 0.1
QM7-X, FR, resampled [†]	5	0.157	5.9 $\pm$ 0.0
	35	1.261	5.8 $\pm$ 0.1
	100	2.442	5.3 $\pm$ 0.2

C.4 Number of distinct training structures

Denoising draws new noise in every epoch, so every training example is a new structure, whereas force regression sees the same displaced structure and force in every epoch. To control for this, we train a *resampled* force-regression model that draws one of the $K = 100$ displaced structures of a molecule uniformly at random each time the molecule is loaded. Everything else, including validation on the fixed displaced structures, is as for force regression. Completion is relative to each run’s own length.

Table 4 counts the distinct structures seen after E epochs over N training molecules. This is NE for denoising, counting every noisy sample as distinct, and N for fixed force regression. In the resampled run a given displaced structure is missed by all E draws with probability $(1 - 1/K)^E$, so the expected number of distinct structures is $NK(1 - (1 - 1/K)^E)$. We report the expected value in the table.

C.5 Noise scale

We pre-train denoising models on QM7-X with $\sigma \in \{0.01, 0.02, 0.04, 0.08, 0.16\}$ Å for 100 epochs without early stopping, otherwise as in appendix A. Each model is fine-tuned from step 95,000 on QM9 U_0 over three seeds. Additionally, we fit linear and gated probes on this pre-training checkpoint.

Results Figure 10 shows the QM9 test error after fine-tuning and the force alignment of the model output for each noise scale. Figures 11 and 12 show the cosine similarity of the linear and equivariant MLP probes to the reference forces across layers.

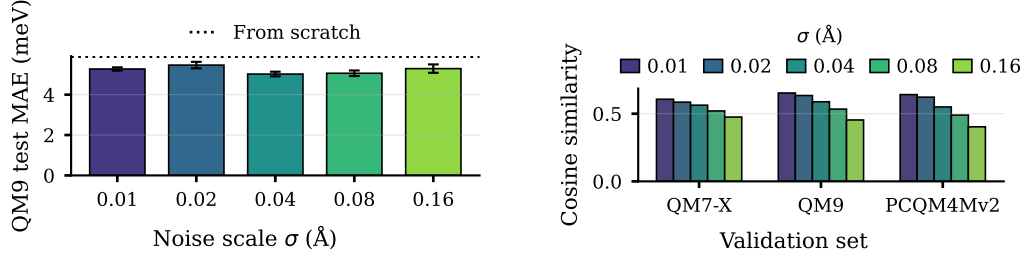


Figure 10: Denoising pre-training on QM7-X at different noise scales σ . Left: QM9 U_0 test MAE after fine-tuning, mean and standard deviation over three seeds. Right: average cosine similarity between the model output and the reference forces of each dataset.

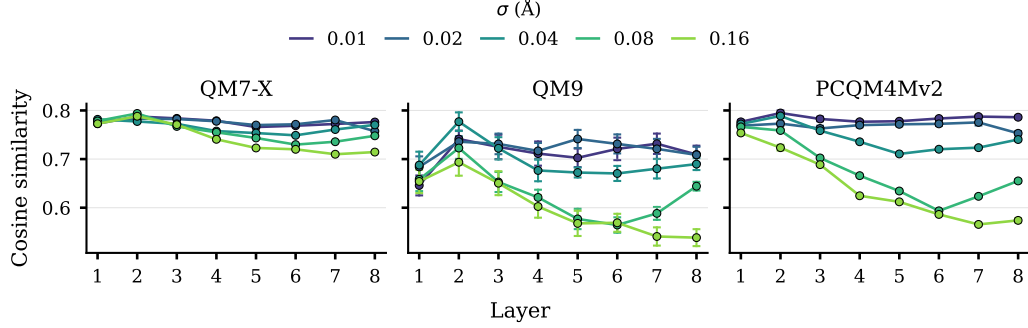


Figure 11: Cosine similarity of a linear probe to the reference forces across layers, for each noise scale σ .

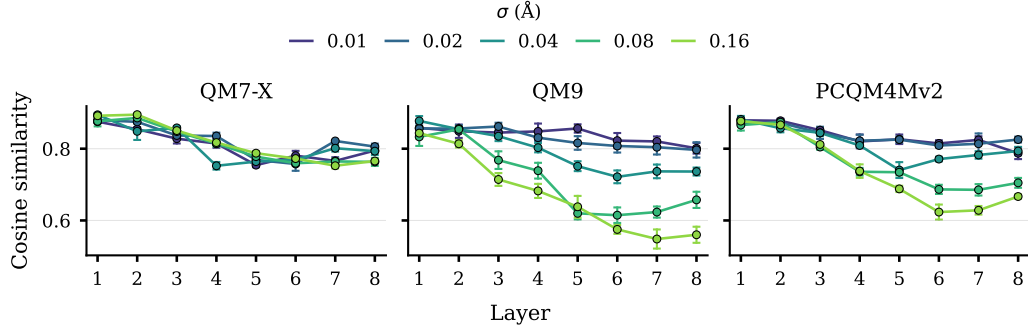


Figure 12: Cosine similarity of an equivariant MLP probe to the reference forces across layers, for each noise scale σ .

D Denoising as supervised force regression

The denoising objective. Zaidi et al. [23] pre-train a model as a denoising auto-encoder. A molecular conformation $\mathbf{x} \in \mathbb{R}^{n \times 3}$ (a local energy minimum) is perturbed with isotropic Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ to give $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$, and a model $r : \mathbb{R}^{n \times 3} \rightarrow \mathbb{R}^{n \times 3}$, the *denoising* network, is trained to predict the noise from the corrupted structure, $\arg \min_{\theta} \mathbb{E} [\|\epsilon - r(\tilde{\mathbf{x}}; \theta)\|_2^2]$. Since the score only matches the force field in the limit $\sigma \rightarrow 0$, predicting the noise is preferred in practice over reconstructing the clean structure. Defining the *reconstruction* network $r_{\mathbf{x}}(\tilde{\mathbf{x}}; \theta) := \tilde{\mathbf{x}} - r(\tilde{\mathbf{x}}; \theta) \approx \mathbf{x}$, the two objectives are equivalent,

$$\|\mathbf{x} - r_{\mathbf{x}}(\tilde{\mathbf{x}}; \theta)\|_2^2 = \|\epsilon - r(\tilde{\mathbf{x}}; \theta)\|_2^2. \quad (1)$$

Vincent [21] has shown that this denoising objective is equivalent to score matching [9] against $q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2 \mathbf{I})$, whose score has the closed form $\nabla_{\tilde{\mathbf{x}}} \log q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x}) = -\epsilon/\sigma^2$. Since $r(\tilde{\mathbf{x}}; \theta)$ is trained to estimate ϵ , its implied score estimate is $-r(\tilde{\mathbf{x}}; \theta)/\sigma^2$, and

$$\|\epsilon - r(\tilde{\mathbf{x}}; \theta)\|_2^2 = \sigma^4 \left\| -\frac{r(\tilde{\mathbf{x}}; \theta)}{\sigma^2} - \nabla_{\tilde{\mathbf{x}}} \log q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x}) \right\|_2^2, \quad (2)$$

so, up to this reparameterization, denoising is equivalent to learning the score of q_{σ} .

Relation to force regression. We show that denoising can be seen as force regression on a second order approximation of the energy function around a conformation $\mathbf{x}$, under certain assumptions on the Hessian.

Theorem 1 (Denoising as force regression). *Let $\mathbf{x} \sim p(\mathbf{x})$, $p(\mathbf{x}) \propto \exp(-U(\mathbf{x}))$, be a molecular conformation at a local energy minimum, i.e. $\nabla_{\mathbf{x}} U(\mathbf{x}) = \mathbf{0}$, and let $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ denote an off-equilibrium structure. Denote by $\mathbf{F}(\tilde{\mathbf{x}}) = -\nabla_{\tilde{\mathbf{x}}} U(\tilde{\mathbf{x}})$ the force acting on $\tilde{\mathbf{x}}$, and let $r(\tilde{\mathbf{x}}; \theta) := \tilde{\mathbf{x}} - r_{\mathbf{x}}(\tilde{\mathbf{x}}; \theta)$ denote the denoising network associated with the reconstruction model $r_{\mathbf{x}}$, as above. Recall that, as shown by Vincent [21], the denoising objective is equivalent to regressing the score of $q_{\sigma}(\tilde{\mathbf{x}} | \mathbf{x})$ via the rescaled estimate $\mathbf{F}_{\theta}(\tilde{\mathbf{x}}) := r(\tilde{\mathbf{x}}; \theta)/\sigma^2$:*

$$\arg \min_{\theta} \mathbb{E} [\|\mathbf{x} - r_{\mathbf{x}}(\tilde{\mathbf{x}}; \theta)\|_2^2] = \arg \min_{\theta} \mathbb{E}_{q_{\sigma}(\tilde{\mathbf{x}}, \mathbf{x})} \left[\left\| \mathbf{F}_{\theta}(\tilde{\mathbf{x}}) - \frac{1}{\sigma^2} (\mathbf{x} - \tilde{\mathbf{x}}) \right\|_2^2 \right]. \quad (3)$$

Then, under the assumption of a locally isotropic Hessian $\nabla_{\mathbf{x}}^2 U(\mathbf{x}) = \mathbf{I}/\sigma^2$, as $\sigma \rightarrow 0$,

$$\arg \min_{\theta} \mathbb{E} [\|\mathbf{x} - r_{\mathbf{x}}(\tilde{\mathbf{x}}; \theta)\|_2^2] = \arg \min_{\theta} \mathbb{E}_{q_{\sigma}(\tilde{\mathbf{x}}, \mathbf{x})} [\|\mathbf{F}_{\theta}(\tilde{\mathbf{x}}) - \mathbf{F}(\tilde{\mathbf{x}})\|_2^2] + o(\sigma^2), \quad (4)$$

i.e. up to the rescaling $\mathbf{F}_{\theta} = r/\sigma^2$, denoising is equivalent to regressing the force acting on the corrupted structure $\tilde{\mathbf{x}}$.

Proof. Using a Taylor expansion of U around $\mathbf{x}$, and $\nabla_{\mathbf{x}} U(\mathbf{x}) = \mathbf{0}$,

$$U(\tilde{\mathbf{x}}) = U(\mathbf{x}) + \frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x})^T \mathbf{H}(\tilde{\mathbf{x}} - \mathbf{x}) + o(\|\tilde{\mathbf{x}} - \mathbf{x}\|^2),$$

where $\mathbf{H} = \nabla_{\mathbf{x}}^2 U(\mathbf{x})$ is the Hessian at $\mathbf{x}$. Differentiating with respect to $\tilde{\mathbf{x}}$, the force at the corrupted structure is

$$\mathbf{F}(\tilde{\mathbf{x}}) = -\nabla_{\tilde{\mathbf{x}}} U(\tilde{\mathbf{x}}) = -\mathbf{H}(\tilde{\mathbf{x}} - \mathbf{x}) + o(\|\tilde{\mathbf{x}} - \mathbf{x}\|).$$

Under the isotropic assumption $\mathbf{H} = \mathbf{I}/\sigma^2$, and using $\tilde{\mathbf{x}} - \mathbf{x} = \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, so that $\mathbb{E}[\|\epsilon\|_2^2] = O(\sigma^2)$,

$$\mathbf{F}(\tilde{\mathbf{x}}) = \frac{1}{\sigma^2} (\mathbf{x} - \tilde{\mathbf{x}}) + o(\sigma^2).$$

Substituting this into the right side of equation 3 gives equation 4, which concludes the proof. $\square$