The Few-Shot Unreliability of Molecular Foundation Models: A Geometric Diagnosis and Partial Remedy

Anonymous Authors[1]

Abstract

Molecular foundation models (MFMs) are widely proposed as a solution to label scarcity in molecular property prediction, on the premise that pretraining on millions of unlabeled molecules produces representations that transfer to new endpoints with minimal supervision. We present an empirical study testing this premise on three MFM models across nine OpenADMET regression endpoints, with training size varying from $N = 10$ to $N = 1000$. We find that frozen MFM embeddings are uniformly worse than Morgan fingerprints at every training size, with the gap widening as the training size grows. A lightweight partial least squares (PLS) projection recovers most of this gap at small training size, revealing that ADMET-relevant directions exist in MFM embeddings but is hidden in directions that a regressor cannot discover from only a handful of labeled examples. Yet even with PLS, the deeper problem remains: at extreme label scarcity, no representation produces R^2 score above a naive train-mean predictor on any OpenADMET endpoint. Specifically, foundation model embeddings do not provide meaningful predictive signal above a naive train-mean baseline until $N \geq 100$, highlighting their unreliability in the label-scarce regimes where they are most needed.

1. Introduction

Predicting molecular properties computationally is one of the central challenges in early drug discovery. Every candidate compound must satisfy a demanding profile of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties before it can advance to clinical trials. A reliable computational property predictor would accelerate drug development timelines by enabling medicinal chemists to prioritize synthesis toward compounds that are more likely to succeed.

One fundamental obstacle is label scarcity. In the molecular domain, each data point requires synthesizing a compound and measuring its properties, a process that can take months to complete. The result is a field that routinely operates on only tens of labeled examples per endpoint, warranting computational methods that can learn from only a few examples.

Molecular Foundation Models (MFMs) (Méndez-Lucio et al., 2024; Shoghi et al., 2024; Feng et al., 2025; Jiang et al., 2025; Wang et al., 2025) have become the dominant proposed answer to this demand. Pretrained on millions of unlabeled molecules, MFMs are expected to produce representations rich enough to transfer to new endpoints with minimal supervision. Yet we find that MFM features consistently underperform classical molecular representations, Morgan fingerprint (FP) (Morgan, 1965; Rogers & Hahn, 2010) and RDKit physicochemical descriptor (Landrum et al., 2026), across nine ADMET endpoints of the OpenADMET benchmark. Furthermore, the performance gap widens rather than closes as more labeled data becomes available.

Using Partial Least Squares (PLS) projection as a diagnostic tool, we show that this failure is geometric rather than informational: ADMET-relevant directions exist within MFM embedding spaces, but are misaligned with the downstream task. A lightweight PLS projection that aligns an MFM's representation recovers most of the performance gap at training size $N = 10$, bringing the prediction performance to effective parity with Morgan fingerprints. Nevertheless, PLS is only a partial remedy. Even with PLS-projection, MFM representations do not outperform a naive train-mean predictor in R^2 score until $N \geq 100$. Our findings underscore the importance of the open question: how to build MFM representations that are aligned with downstream tasks from the outset, particularly when labeled data is scarce.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the AI for Science workshop (ICML 2026). Do not distribute.

2. Background

2.1. Molecular Foundation Models

Modern molecular foundation models follow the pretraining-then-finetuning paradigm. Uni-Mol (Zhou et al., 2023) employs a 3D-aware transformer pretrained on millions of molecular conformers via 3D structure reconstruction and masked atom prediction. MolFormer (Wu et al., 2023) uses a linear-attention transformer with rotary embeddings, pretrained with the masked language modeling approach. ChemBERTa (Chithrananda et al., 2020) adapts the RoBERTa (Liu et al., 2019) architecture to SMILES (Weininger, 1988). All three models produce fixed-dimensional embeddings that can be used as features for downstream supervised predictors.

2.2. Classical Representations

Morgan fingerprints (Morgan, 1965; Rogers & Hahn, 2010), also known as extended-connectivity fingerprints (ECFP), encode local circular chemical environments into sparse binary vectors. Despite their conceptual simplicity, they remain competitive across many ADMET prediction tasks. RDKit physicochemical descriptors (Landrum et al., 2026) encode domain-expert knowledge about drug-likeness (Lipinski properties, TPSA, logP) and functional group composition, and have been popular for QSAR modeling for decades.

2.3. Few-Shot Molecular Property Prediction

Quantitative structure-activity relationship (QSAR) modeling has long sought to predict molecular properties from chemical structure, dating to the foundational work ofHansch & Fujita (1964) and the development of interpretable descriptor-based regressors such as partial least squares (Wold et al., 2001). These classical pipelines pair hand-crafted molecular representations with classical regression models, and remain surprisingly competitive in low-data regimes (Van Tilborg et al., 2022). The introduction of message-passing neural networks (Gilmer et al., 2017) and pre-trained graph transformers (Hu et al., 2019) shifted the community toward deep molecular representations, with benchmarks such as MoleculeNet (Wu et al., 2018) and TDC (Huang et al., 2021) establishing standard evaluation protocols. However, these benchmarks predominantly assess performance at a relatively large training set size of hundreds to thousands of labeled molecules, obscuring behavior in the few-shot regime.

3. Partial Least Squares as a Geometric Probe and Remedy

In the past, when QSAR modeling faced datasets with only dozens of compounds but hundreds of molecular descriptors, Wold et al. (1984) introduced Partial Least Squares regression as a principled response. Rather than operating in the full descriptor space, they operate on the low-dimensional subspace that exhibit maximal covariance with the biological response.

Foundation model embeddings have recreated the same problem at larger scale. MFMs often produce continuous representations with hundreds or thousands of dimensions, pretrained to capture global molecular diversity. With only $N = 10$ labeled examples, a supervised learner has no ability to identify which directions of this space are relevant to a specific ADMET endpoint.

Within an MFM's embedding space, partial least squares (PLS) can find the right directions that maximizes embeddings-labels covariance. Concretely, given a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, PLS finds a sequence of k vectors $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_k \,|\boldsymbol{w}_j \in \mathbb{R}^d\}$ that successively maximize the covariance between the projected scalar $t_j = \boldsymbol{x} \cdot \boldsymbol{w}_j$ and the response y, deflating the variance explained by previous components at each step. A downstream predictor (e.g., a random forest) can then learn to predict the ADMET properties from the resulting k-dimensional score matrix $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W} \in \mathbb{R}^{n \times k}$.

Our use of PLS departs from the classical QSAR tradition in one critical respect. In classical QSAR, PLS is the regressor: predictions are made as $\hat{y} = \boldsymbol{T}\boldsymbol{q}^T + b$, where $\boldsymbol{q} \in \mathbb{R}^k$ is a learned weight vector that maps the k latent components to the predicted output. We decouple the projection from the regression entirely, feeding the projected features $\boldsymbol{T}$ into a separate nonlinear predictor that can exploit structure in the aligned latent space that a linear regressor would miss. This also means PLS serves as a diagnostic tool: if the projection improves performance substantially, the embedding contained task-relevant information all along, just in the wrong directions. If it does not, the information was simply absent. The size of the improvement after PLS-projections is therefore also a direct measure of geometric misalignment.

4. Experiments

4.1. Foundation Model Embeddings Consistently Underperform Classical Representations

Setup. We train random forest (RF) regressor heads on nine regression datasets from OpenADMET (Fraser et al., 2025) using five representation types: Morgan fingerprints (baseline) (Morgan, 1965; Rogers & Hahn, 2010), RDKit

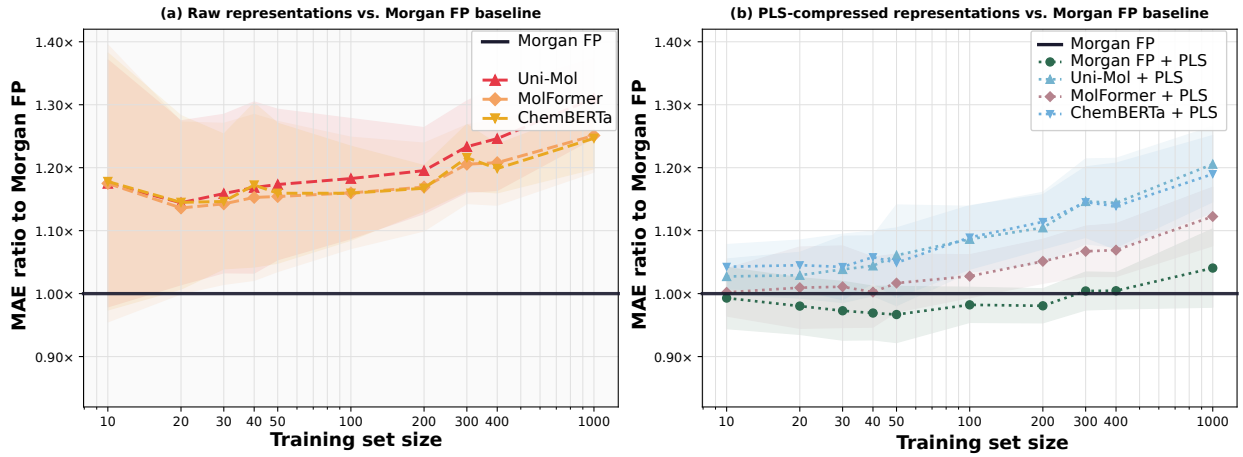

Figure 1. MAE ratio relative to the Morgan FP regressor across training sizes (lower is better; 1.0 = parity with baseline). (a) Raw MFM embeddings are uniformly worse than Morgan FP at every training size. (b) PLS projection ($k = 10$) recovers most of the gap at small N. PLS-projected MolFormer reaches effective parity at $N = 10$, but erodes as N reached a threshold.

physicochemical descriptors (Landrum et al., 2026), Uni-Mol (Zhou et al., 2023), MolFormer (Wu et al., 2023), and ChemBERTa (Chithrananda et al., 2020). For all representations, we use the same RF regressor head; MFM models are frozen and used purely as feature encoders, ensuring that differences in performance reflect the representation rather than learning capacity. We use RF throughout the experiments as it consistently outperformed Gaussian Process regressors and linear probes across all training sizes and endpoints in preliminary experiments, giving MFM embeddings the most favorable evaluation conditions and ensuring the reported gaps are conservative. We vary training size to $N = \min(|\mathbb{D}|, n)$ where $\mathbb{D}$ is the full dataset and $n \in \{10, 20, 30, 40, 50, 100, 200, 300, 400, 1000\}$. We report mean absolute error (MAE) averaged over 20 random seeds as a ratio relative to the Morgan FP baseline. A ratio greater than 1.0 indicates the representation is worse than Morgan fingerprints.

Results. Figure 1(a) shows that all three foundation models exceed a ratio of 1.0 at every training size tested. At $N = 10$, Uni-Mol, MolFormer, and ChemBERTa each produce a 17–18% performance deficit relative to Morgan FP. The gap does not close as data accumulates: by $N = 1000$, it has widened to 25–31%, indicating that Morgan fingerprints scale better with additional labeled data than high-dimensional continuous embeddings do.

A natural concern with multi-endpoint benchmarks is that a single hard endpoint can dominate the arithmetic mean and create apparent method differences that vanish under normalized comparison. Figure 2 rules this out. Panel (a) shows mean rank across endpoints: the three raw MFM embeddings (Uni-Mol, MolFormer, ChemBERTa) consistently rank at 7–9 across the entire learning curve, never approaching the top half of the ranking. Panels (b) and (c) show per-endpoint MAE ratios at $N = 10$ and $N = 200$, respectively. While the spread across endpoints is large, the *ordering* of methods is nearly identical at both training sizes, and the mean (diamond) for each MFM family sits well above 1.0 at both. The underperformance of MFM embeddings is thus consistent and endpoint-agnostic.

4.2. PLS Projection Diagnoses a Geometric Misalignment

Setup. We apply PLS as a supervised dimensionality reduction step prior to RF regression, projecting each embedding onto the $k = 10$ directions of maximum label–feature covariance estimated from the training set. PLS has been used as a regressor in QSAR (explained in Section 2), but here we use it strictly as a representation alignment step, before the final nonlinear predictor is applied. If MFM embeddings simply lacked task-relevant information, PLS would provide no benefit; it can only recover directions that exist. The size of the improvement is therefore a direct measure of how much relevant information is geometrically hidden in the raw embedding.

Results. Figure 1(b) shows that, at $N = 10$, PLS projection brings all three MFM families close to Morgan FP parity. MolFormer's MAE ratio drops from 1.175 to 1.002, effectively eliminating the 17.5% performance gap. Similar trends appear for Uni-Mol and ChemBERTa. Critically, the same projection applied to Morgan fingerprints yields only a 0.7% improvement, consistent with PLS providing greater benefit to high-dimensional continuous embeddings than to sparse binary ones.

The near-total recovery at $N = 10$ implies that ADMET-relevant information, up to what is captured by Morgan FP, is present in MFM embeddings but is hidden in directions

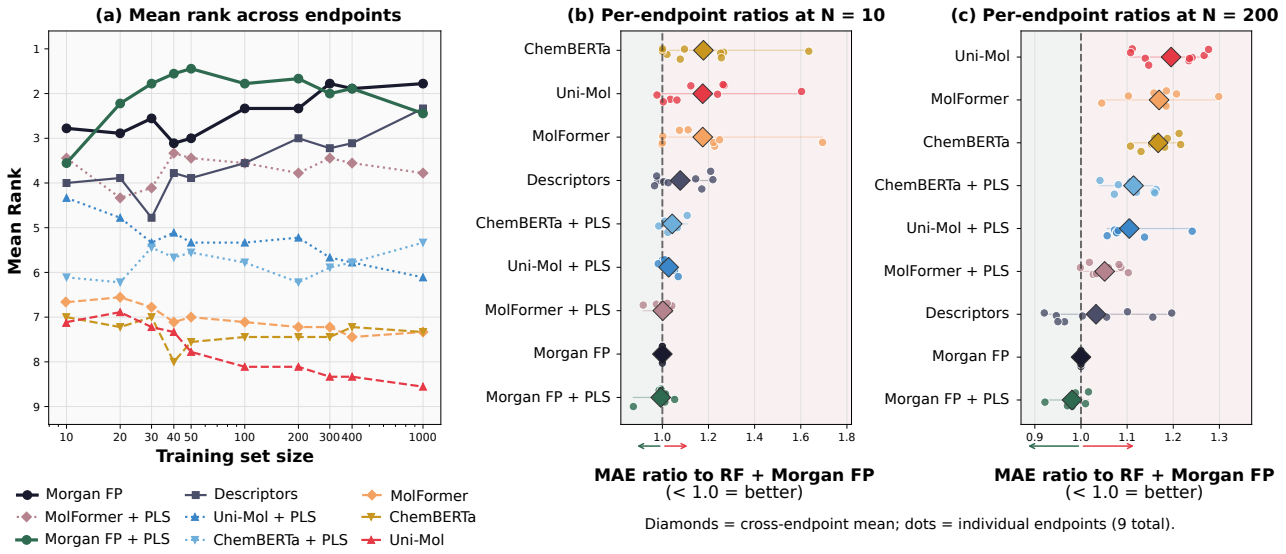

Figure 2. Method ranking is stable across all nine endpoints and is not driven by any single hard endpoint. (a) Raw MFM embeddings rank $7-9$ across the full learning curve; Morgan FP variants hold ranks 1–2. (b, c) Per-endpoint MAE ratios at $N = 10$ and $N = 200$. Despite wide endpoint-level spread (dots), cross-endpoint means (diamonds) are ordered nearly identically at both training sizes.

that are mostly orthogonal to the top principal components. This is geometrically expected: Uni-Mol, MolFormer, and ChemBERTa are pretrained on general chemical corpora with objectives that reward global molecular fidelity. The principal axes of their embedding spaces therefore capture broad chemical diversity, molecular size, aromaticity, scaffold class, not the local functional group variation that governs ADMET properties. A Random Forest operating in this 512–768-dimensional space with only 10 labeled examples has no ability to discover the relevant directions on its own.

By $N = 200$, PLS-compressed MFM features have rewidened to ratios of 1.05–1.11, and their advantage over raw embeddings narrows. This is not a failure of PLS, as PLS-projected MFMs still outperform raw MFMs. Instead, as labeled data accumulates, an RF regressor can begin to identify ADMET-relevant directions in the raw embedding directly, making the PLS pre-alignment progressively less critical. The baseline, however, scales better still: Morgan fingerprints, whose sparse binary structure provides direct substructure indexing, extract more signal from each additional label than any high-dimensional continuous embedding does.

PLS is thus a diagnostic as much as a remedy. It establishes that MFM underperformance in the few-shot regime is a fixable geometric problem, not an absence of relevant information. But the fix is inherently transient: it is most valuable precisely where labeled data is most scarce.

4.3. All Methods Fail below the Threshold that Matters

While we have established that MFMs underperform Morgan fingerprints, and that PLS partially restores them, the metric is still relative to other methods. A more fundamental question for practitioners is whether any of these methods provides *useful predictive signal* for drug discovery in real-world. To answer this, we shift from MAE ratio to R^2. As baselines, we use the test-mean oracle ($R^2 = 0$), a hypothetical predictor that always outputs the test set mean, and the train-mean oracle that always outputs the train set mean.

Figure 3 shows that at $N = 10$, the mean R^2 of every representation falls between or below the train-test oracles R^2 lines. The RF regressor with Morgan FP achieves the best mean R^2, but still at similar level as a train-mean oracle. Meanwhile, raw MFM-based regressors sit at $R^2 = -0.25$ to -0.16, below even the naive train-mean oracle. PLS-compressed MolFormer, the best MFM variant, reaches -0.09, marginally above the train-mean baseline. These results suggest that with 10 labeled molecules, no representation provides substantially more predictive power than a naive predicting-train-mean strategy can achieve.

In our sweep, classical representations begin to provide meaningful benefits at $N = 30$, while MFM representations begin at $N = 100$. Specifically, Morgan FP crosses the test-mean oracle line at $N = 30$, followed closely by descriptors. PLS-projected MolFormer is the first MFM to cross the test-mean oracle line, at $N = 100$. At $N = 200$, all methods show positive mean R^2, and genuine predictive power is present.

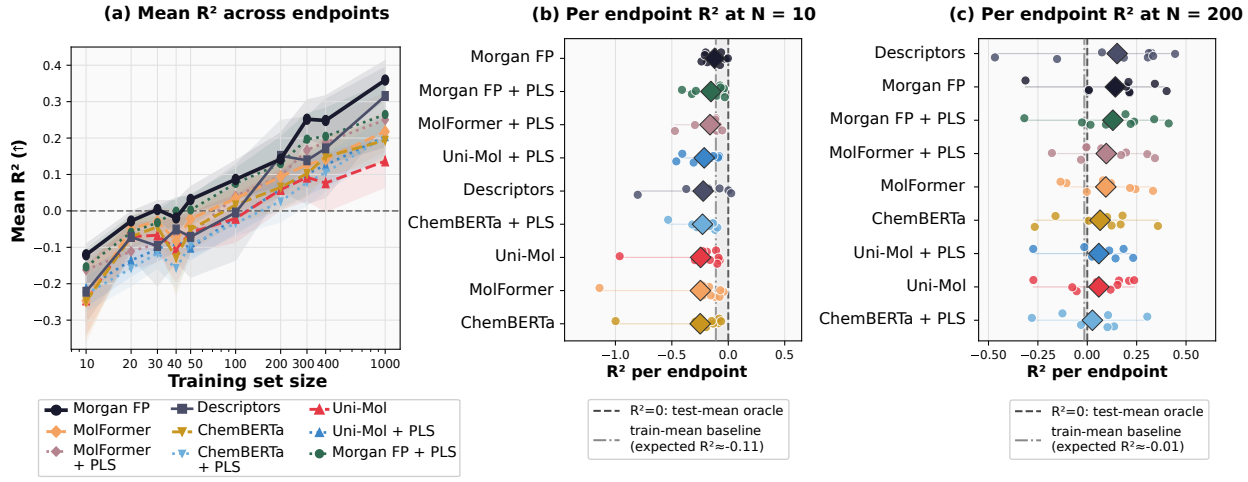

Figure 3. (a) Mean R^2 scores across all datasets at various training set size. (b) At $N = 10$, no representation clearly outperform the train-mean oracle ($R^2 = -0.11$), meaning no method provides actionable predictive power. (c) At $N = 200$, all methods show positive R^2 scores.

PLS as a geometric fix narrows the gap between MFMs and Morgan FP, but without improving the usefulness of their predictions against a simple train-mean oracle. In the label-scarce regime that matters most for early drug discovery, the right question is not which representation wins, but rather whether the labeled data budget is sufficient for any data-driven model to function at all. For ADMET data, our results suggest that budget is approximately $N \geq 30$ for classical representations and $N \geq 100$ for foundation models, thresholds that should inform experimental design and method choice.

5. Limitations

Our study uses foundation model embeddings in a transfer (frozen) setting without fine-tuning. It is possible that fine-tuning on the few-shot training set, if done with appropriate regularization, recovers some performance. However, fine-tuning with $N = 10 - 50$ can be very unstable in practice. We also note that our conclusions apply to the specific OpenADMET endpoints studied; different chemical spaces or task types may show different patterns. Finally, we evaluate only regression tasks; classification ADMET tasks may behave differently.

6. Conclusion

We have demonstrated that MFM features underperform classical molecular features in the few-shot regime. The few-shot unreliability of MFMs is, at its core, a misalignment between what pretraining optimizes and what drug discovery requires. Pretraining on millions of structurally diverse molecules produces embeddings whose geometry reflects global chemical diversity, while ADMET prediction in the few-shot regime demands sensitivity to local functional group variation that occupies a small corner of that learned space. Partial least square (PLS) projection offers a post-hoc remedy that partially aligns the features to the task. However, a post-hoc projection is not a substitute for an objective that targets the right geometry from the start. Redesigning pretraining strategies around the statistical structure of property-relevant chemical variation, rather than around molecular reconstruction or contrastive diversity, is a promising direction we leave as the natural next step of this ongoing work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Feng, S., Ni, Y., Ma, Z.-M., Ma, W.-Y., Lan, Y., et al. Unigem: A unified approach to generation and property prediction for molecules. In *The Thirteenth International Conference on Learning Representations*, 2025.

Fraser, J. S., Edgar, S., Handly, L. N., Kosuri, S., Chodera, J. D., Murcko, M., and Walters, W. P. Openadmet: Embracing the avoid-ome to transform drug discovery. 2025.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chem-

istry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Hansch, C. and Fujita, T. p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8):1616–1626, 1964.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Jiang, F., Prakash, M., Ma, H., Deng, J., Guo, Y., Mollaysa, A., Mansi, T., Liao, R., and Huang, J. Trident: Tri-modal molecular representation learning with taxonomic annotations and local correspondence. *Advances in Neural Information Processing Systems*, 2025.

Landrum, G. et al. Rdkit: Open-source cheminformatics software, 2026. URL https://www.rdkit.org.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Méndez-Lucio, O., Nicolaou, C. A., and Earnshaw, B. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Shoghi, N., Kolluru, A., Kitchin, J. R., Ulissi, Z. W., Zitnick, C. L., and Wood, B. M. From molecules to materials: Pre-training large generalizable models for atomic property prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Van Tilborg, D., Alenicheva, A., and Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23):5938–5951, 2022.

Wang, L., Liu, S., Rong, Y., Zhao, D., Liu, Q., and Wu, S. Molspectra: Pre-training 3d molecular representation with multi-modal energy spectra. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

Wold, S., Sjöström, M., and Eriksson, L. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

Wu, F., Radev, D., and Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5312–5320, 2023.

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.

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.