MOLFIBER: FROM GLOBAL ANCHORS TO LOCAL REFINEMENT IN MOLECULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular property prediction increasingly relies on combining complementary representations, yet standard fusion can be ineffective when one representation is already strong. In this regime, the key question is not how to combine all views globally, but where additional representations provide information that the strong representation misses.

We introduce **MolFiber**, a Morgan-anchored framework that treats Morgan fingerprints as a global reference geometry for chemical space and uses GraphGrid and MolFormer as local refiners within neighborhoods of Morgan-similar molecules. Rather than acting as competing global predictors, the auxiliary views resolve distinctions that Morgan cannot see locally. **MolFiber-G** uses the full fiber-conditioned auxiliary evidence, whereas **MolFiber-L** uses only locally centered contrast; their ensemble, **MolFiber-ENS**, combines these complementary inductive biases. Across 24 molecular property prediction benchmarks, MolFiber-ENS achieves the strongest aggregate performance among the compared methods and outperforms controlled global-fusion baselines using the same representations. These results support a broader principle: complementary representations can be most valuable not when fused globally, but when used to refine the local ambiguities of a strong anchor.

1 INTRODUCTION

Molecular property prediction depends strongly on molecular representation. Morgan fingerprints encode local chemical substructures; graph models capture connectivity and organization; pretrained molecular language models learn contextual structure from large chemical corpora; and 3D models encode conformation (Rogers & Hahn, 2010; Gilmer et al., 2017; Ross et al., 2022; Zhou et al., 2023). These complementary views make multirepresentation learning appealing, yet simply combining them often yields limited gains (Praski et al., 2025). Concatenation, attention, and late fusion must reconcile representations that live in different spaces, operate at different scales, and may be informative in different regions of chemical space.

The problem is especially interesting when one representation is already strong. Morgan fingerprints remain remarkably competitive with modern learned representations on molecular property prediction (Rogers & Hahn, 2010; Praski et al., 2025). In this regime, an auxiliary representation need not outperform Morgan globally to be useful; it only needs to reveal information that Morgan cannot see.

This suggests a different question from conventional fusion: *where does a strong representation become locally ambiguous, and can another representation resolve that ambiguity?*

Our perspective is geometric. Morgan–Tanimoto similarity provides a *reference geometry* for chemical space: molecules sharing many local substructures lie nearby. A Morgan predictor then defines a prediction landscape over this space. Much of that landscape may already be well resolved; the interesting regions are those where Morgan-nearby molecules nevertheless differ in property. Such molecules occupy nearly the same location according to the anchor, yet may differ in graph organization, longer-range structure, or contextual patterns visible to another representation.

We introduce **MolFiber**, a coarse-to-fine framework built around this idea. Morgan plays two roles: it provides a strong global predictor and defines local neighborhoods of Morgan-similar molecules, which we call *approximate Morgan fibers*. GraphGrid and MolFormer (Ross et al., 2022) then act as

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

local lenses: rather than forming competing global predictors, they compare molecules within these fibers and learn distinctions that Morgan compresses or cannot represent. Morgan therefore charts the global space, while the auxiliary views provide finer resolution where that map becomes locally ambiguous.

For property t, MolFiber retains the frozen Morgan prediction and adds only a bounded, neighborhood-conditioned correction,

$$S^{(t)}(x) = b_M^{(t)}(x) + \Delta^{(t)}(x), \qquad |\Delta^{(t)}(x)| \le \lambda.$$

The auxiliary views therefore refine the anchor without an unrestricted path to replace its global prediction scale.

The fiber-conditioned auxiliary score naturally decomposes as

$$u_{x,x} = \underbrace{u_{x,x} - \bar{u}_x}_{\text{local contrast}} + \underbrace{\bar{u}_x}_{\text{fiber-level signal}}.$$

This decomposition gives two complementary ways to refine Morgan. **MolFiber-G** retains the full contextual score $u_{x,x}$, preserving both molecule-specific contrast and information shared across the local fiber. **MolFiber-L** centers the score within the fiber and uses only $u_{x,x} - \bar{u}_x$, isolating what distinguishes the query from molecules that Morgan regards as similar. The former preserves potentially useful regional information; the latter removes neighborhood-level shifts and focuses purely on relative local evidence.

Rather than forcing one assumption to hold everywhere, we also introduce **MolFiber-ENS**, which independently trains MolFiber-G and MolFiber-L and averages their prediction logits. The two models encode complementary inductive biases—full fiber-conditioned evidence versus locally centered contrast—and their ensemble provides the strongest aggregate performance in our experiments. We additionally study learned task-level, molecule-adaptive, and routing variants in the appendix; notably, the simpler G/L construction and its ensemble remain stronger overall.

Crucially, MolFiber is more than another fusion architecture. We compare against controlled fusion baselines given the same Morgan, GraphGrid, and MolFormer representations, including a model that preserves the frozen Morgan anchor and bounded residual but removes the fiber construction. MolFiber consistently improves over these controls, indicating that the gain comes not simply from adding representations, but from organizing them asymmetrically: a strong representation defines the global geometry, while complementary representations resolve its local ambiguities.

We evaluate MolFiber on 24 molecular property prediction benchmarks from MoleculeNet (Wu et al., 2018) and Therapeutics Data Commons (Huang et al., 2021) against thirteen representative fingerprint, graph, pretrained, and hybrid baselines. MolFiber-G is highly competitive as a single model, while MolFiber-ENS achieves the strongest aggregate performance across the full benchmark suite. Controlled fusion experiments and architectural ablations further isolate the benefit of Morgan-local conditioning. Together, these results support a broader principle: *a representation need not be globally strong to be useful; its value may emerge precisely where a stronger representation loses local resolution.*

Our main contributions are:

- We introduce an *anchor-relative* approach to molecular multirepresentation learning, in which a strong representation defines the global prediction geometry and complementary views resolve ambiguity within anchor-defined neighborhoods.
- We formulate *approximate Morgan fibers* and decompose local auxiliary evidence into molecule-specific contrast and fiber-level signal, providing a simple coarse-to-fine view of multirepresentation learning.
- We propose **MolFiber-G**, which retains full fiber-conditioned evidence, **MolFiber-L**, which uses locally centered contrast, and **MolFiber-ENS**, which combines their complementary inductive biases.
- Through controlled fusion and architectural ablations across molecular benchmarks, we show that Morgan-local organization is more effective than using the same auxiliary representations as unrestricted global features, with MolFiber-ENS achieving the strongest aggregate performance among the compared methods.

2 Related Work

Molecular Representations for Property Prediction. Molecular property prediction spans a wide range of representations, with no family uniformly dominant. Extended-connectivity (Morgan) fingerprints (Rogers & Hahn, 2010) encode each atom's circular substructure environment and remain a fast, deterministic, learning-free baseline. Graph neural networks and graph-transformer architectures (Gilmer et al., 2017; Ying et al., 2021; Maziarka et al., 2020; 2024) learn directly from molecular connectivity, sometimes augmenting attention with inter-atomic distance or bond structure for chemical inductive bias. A third line pretrains transformer language models on unlabeled SMILES corpora (Chithrananda et al., 2020; Fabian et al., 2020; Ross et al., 2022; Winter et al., 2019); a fourth pretrains on 3D conformations (Zhou et al., 2023; Ji et al., 2024) or via self-supervised graph objectives (Rong et al., 2020; Liu et al., 2022), injecting information from far more molecules than any single downstream task provides.

Despite this diversity, large-scale, fair comparisons find plain Morgan/ECFP fingerprints remain difficult to outperform (Praski et al., 2025): message-passing GNN embeddings are frequently of markedly lower quality than fingerprint or pretrained-transformer representations, and models such as MolFormer show negligible aggregate improvement over ECFP. This is the backdrop for our design: rather than treating Morgan as one view to fuse, we treat it as the anchor every auxiliary representation must be measured against.

Combining Molecular Representations. A large body of work combines multiple molecular representations (fingerprints, graphs, SMILES, 3D geometry) into a single predictor via concatenation, attention, or learned weighting (Wang et al., 2022), and contrastive fingerprint-language fusion pushes this further by aligning molecular and textual embeddings (Seidl et al., 2023). These approaches share a structure: every representation gets an unrestricted path to a global scoring function, and the model must learn from data when each view is trustworthy. This suits comparably strong views, not the regime above, where Morgan is already hard to beat globally and the rest contribute unevenly across chemical space. Fusing a weak global signal into a strong one risks diluting the predictor rather than improving it. Our departure is structural, not architectural: GraphGrid and MolFormer never receive a global prediction head in MolFiber, entering only through comparisons localized to a Morgan-defined neighborhood, made precise in section 3.

Locality, Relative Comparison, and Activity Cliffs. That structurally similar molecules can differ sharply in property is established in medicinal chemistry as *activity cliffs* (Stumpfe & Bajorath, 2012): compounds a similarity measure places close together whose activities nonetheless diverge, precisely the failure mode approximate Morgan fibers isolate. Computational work exploits relative comparison directly: matched molecular pair analysis and pairwise delta-property models (Fralish et al., 2023) predict property *differences* rather than absolute values, and few-shot approaches condition on a support set of reference molecules (Schimunek et al., 2023). These share our intuition that relative comparisons are often more reliable than absolute predictions alone.

Positioning. MolFiber differs in how the comparison is anchored. Matched-pair and few-shot-context methods define similarity and support from the auxiliary representation itself, with no independently strong global predictor as fallback. MolFiber instead applies relative comparison *asymmetrically*: Morgan alone defines both the global coordinate system and the neighborhoods, while GraphGrid and MolFormer contribute only bounded, Morgan-local corrections. This keeps the global scale well-defined through the anchor, while letting the auxiliary views express activity-cliff-like signal that a global fusion model would drown out or miscalibrate. MolFiber thus sits between standard multiview fusion, which grants every representation unrestricted global reach, and pure relative-comparison methods, which forgo a strong global anchor altogether.

3 Methodology

Overview and Motivation. MolFiber starts from a simple premise: when one molecular representation is already strong, additional representations need not compete with it globally. Instead, they can focus on what the strong representation cannot resolve: *among molecules that look nearly identical to the anchor, can another view tell them apart?* Such local ambiguities include the classical *activity-cliff* regime (Stumpfe & Bajorath, 2012), in which structurally similar molecules exhibit

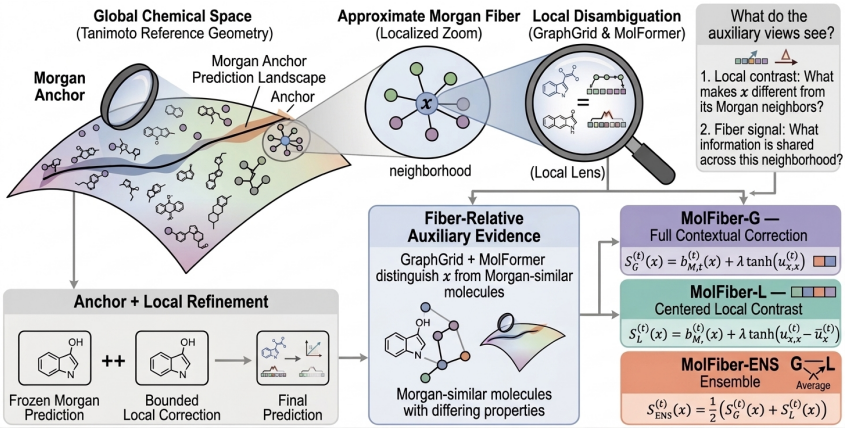

Figure 1: **MolFiber: coarse-to-fine local refinement of a strong molecular anchor.** Morgan fingerprints define the global reference geometry, while a frozen Morgan predictor provides the anchor prediction. Morgan–Tanimoto neighborhoods define local fibers of structurally similar molecules. GraphGrid and MolFormer then act as local lenses that reveal distinctions compressed by the Morgan representation. Their evidence separates naturally into molecule-specific local contrast and information shared across the fiber. MolFiber-G retains both components, whereas MolFiber-L keeps only the centered local contrast. MolFiber-ENS combines the two complementary predictions. The detailed computational architecture is given in Appendix Figure 2.

sharply different activities, but the same principle applies more broadly to molecular properties. Figure 1 illustrates this coarse-to-fine principle.

We use three deliberately different molecular views: a Morgan fingerprint $M(x)$, a GraphGrid representation $G(x)$ of molecular graph organization, and a pretrained MolFormer embedding $F(x)$ (See App. A.1). Their roles are asymmetric. Morgan acts as the *anchor*: it provides a global property prediction and, through Tanimoto similarity, a reference geometry for chemical space. Nearby molecules in this geometry define an *approximate Morgan fiber*. GraphGrid and MolFormer then act as local lenses within each fiber, resolving distinctions that Morgan compresses or cannot represent.

Thus, Morgan tells the model *where the molecule lies*, while the auxiliary views help determine *what Morgan cannot distinguish there*. They are never given an unrestricted path to a competing global predictor. For target t, every MolFiber prediction takes the form

$$S^{(t)}(x) = b_M^{(t)}(x) + \Delta^{(t)}(x), \qquad \left|\Delta^{(t)}(x)\right| \le \lambda,$$

where $b_M^{(t)}(x)$ is the frozen Morgan anchor and $\Delta^{(t)}(x)$ is a bounded, neighborhood-conditioned correction. The complete computational pipeline and implementation details are given in Appendix A and Figure 2.

3.1 Morgan Anchors and Approximate Fibers

We first fit a random forest f_M on Morgan fingerprints and convert its positive-class probabilities to logits,

$$b_M(x) = \mathrm{logit}\big(f_M(M(x))\big) \in \mathbb{R}^T.$$

The Morgan predictor is then frozen.

Morgan also defines the geometry in which local ambiguity is measured. Let

$$s_M(x, z) = \mathrm{Tan}\big(M(x), M(z)\big)$$

denote Morgan–Tanimoto similarity. For each molecule x, we retrieve its k nearest training molecules,

$$\mathcal{F}_k(x) = \mathrm{kNN}_M\left(x; \mathcal{D}_{\mathrm{train}}\right), \qquad \mathcal{B}_x = \{x\} \cup \mathcal{F}_k(x).$$

We call $\mathcal{B}_x$ an *approximate Morgan fiber*: a local region containing molecules that Morgan places near the same location in chemical space.

These are precisely the regions where complementary representations can be most useful. Molecules may share many Morgan substructures yet differ in graph organization or contextual structure captured

by GraphGrid and MolFormer. MolFiber therefore uses the auxiliary views to resolve variation *within* Morgan fibers rather than to replace Morgan globally. At inference, neighbors are retrieved only from the training set. The out-of-fold construction of training anchors, fiber construction, and leakage controls are detailed in Appendix A.

3.2 Fiber-Relative Auxiliary Evidence

GraphGrid and MolFormer are first mapped to a common latent width,

$$g_x = P_G(G(x)), \qquad m_x = P_F(F(x)),$$

where P_G is a lightweight GraphGrid encoder and P_F projects the frozen MolFormer embedding. Crucially, neither representation is fed directly to a global property head. Auxiliary information enters only through comparisons between molecules in the same Morgan fiber.

For each ordered pair $z, w \in \mathcal{B}_x$, we construct

$$r_{zw} = \big[g_z - g_w,\ g_z \odot g_w,\ m_z - m_w,\ m_z \odot m_w,\ s_M(z,w)\big], \qquad e_{zw} = \varphi_\theta(r_{zw}).$$

The difference terms describe how the auxiliary views distinguish two Morgan-similar molecules, the product terms capture what they share, and $s_M(z,w)$ supplies the strength of their anchor-level similarity.

Each $z \in \mathcal{B}_x$ aggregates its comparisons with the other fiber members to produce a target-wise contextual score

$$u_{x,z} \in \mathbb{R}^T.$$

The exact similarity-weighted aggregation is given in Appendix A.2.

For the query molecule, we compare its contextual score $u_{x,x}$ with the Morgan-weighted neighborhood mean

$$\bar{u}_x = \sum_{z \in \mathcal{F}_k(x)} \beta_{xz} u_{x,z}.$$

This gives the decomposition

$$u_{x,x} = \underbrace{u_{x,x} - \bar{u}_x}_{\text{local contrast}} + \underbrace{\bar{u}_x}_{\text{fiber-level signal}}.$$

The first term captures what distinguishes x from molecules Morgan regards as similar, while the second captures auxiliary evidence shared across the local fiber.

3.3 MolFiber-G, MolFiber-L, and MolFiber-ENS

The decomposition above naturally defines a one-parameter family of corrections,

$$S_\rho^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\Big(u_{x,x}^{(t)} - (1-\rho)\,\bar{u}_x^{(t)}\Big), \qquad \rho \in [0,1], \tag{1}$$

where ρ controls how much of the fiber-level signal is retained alongside the local contrast. In the main method, we study the two endpoints $\rho = 1$ and $\rho = 0$; Appendix D.2 studies learned intermediate values.

MolFiber-G ($\rho = 1$): full contextual evidence. MolFiber-G retains the complete contextual score,

$$S_G^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\big(u_{x,x}^{(t)}\big). \tag{2}$$

It uses both the molecule-specific local contrast and the signal shared by its Morgan neighborhood, at the cost of also inheriting any bias shared across the fiber — for instance, if the auxiliary encoders are systematically miscalibrated in that region of chemical space.

MolFiber-L ($\rho = 0$): local contrast. MolFiber-L centers the contextual score within the fiber,

$$S_L^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\big(u_{x,x}^{(t)} - \bar{u}_x^{(t)}\big). \tag{3}$$

Centering removes any component shared across the neighborhood and retains only how the query differs from its Morgan-similar neighbors. This makes the correction insensitive to a common shift in auxiliary evidence within a fiber, at the cost of discarding potentially predictive fiber-level signal.

MolFiber-G and MolFiber-L therefore encode complementary inductive biases at the two ends of Equation 1: the former preserves absolute fiber-conditioned evidence, while the latter uses a purely relative local correction. Appendix D.2 studies two learned alternatives to fixing ρ at an endpoint outright, together with a routing variant.

MolFiber-ENS. Rather than fixing a single value of ρ within one model, MolFiber-ENS independently trains MolFiber-G and MolFiber-L and averages their logits,

$$S_{\mathrm{ENS}}^{(t)}(x) = \frac{1}{2}\left(S_G^{(t)}(x) + S_L^{(t)}(x)\right). \tag{4}$$

The ensemble combines fiber-level and locally centered evidence through two separately optimized predictors and is our strongest aggregate model.

All variants are trained with masked binary cross-entropy. The contextual readout is zero-initialized, so training starts exactly from the frozen Morgan anchor, while the bounded residual guarantees

$$\left|S^{(t)}(x) - b_M^{(t)}(x)\right| \le \lambda.$$

Full training, optimization, leakage-control, and computational details are given in Appendix A.

Beyond the three variants studied here, Appendix D.2 reports two learned alternatives to fixing ρ at an endpoint — MolFiber-M, which learns a single task-level $\rho^{(t)}$, and MolFiber-A, which estimates $\rho^{(t)}(x)$ adaptively per molecule — together with MolFiber-route, an architectural ablation that routes between GraphGrid and MolFormer as separate relative experts rather than combining them jointly. Neither learned variant nor the routing ablation exceeds MolFiber-G or MolFiber-ENS in aggregate performance on our benchmark suite.

4 Experiments

We evaluate MolFiber on molecular property prediction (MPP). Hyperparameters, including k, λ, temperatures, latent dimensions, and loss weights, are selected exclusively on held-out validation splits disjoint from the reported test sets; no test-target specific tuning is performed.

Datasets. We evaluate on 24 molecular property prediction benchmarks: six classification tasks from MoleculeNet (Wu et al., 2018) and 18 ADME/toxicity endpoints from the Therapeutics Data Commons (Huang et al., 2021), spanning single- and multi-task binary classification under the scaffold-based splits associated with each source. Three datasets are multitask, handled with the masked formulation of section A.3. Dataset names, sizes, task counts, and label statistics appear in section C.1.

Pretrained and fingerprint baselines. We compare against 13 representative molecular representation models spanning fingerprint, fingerprint-fusion, graph-transformer, SMILES/text- transformer, 3D-conformer, hybrid GNN-transformer, and GNN self-supervised-pretraining paradigms: CLAMP (Seidl et al., 2023), R-MAT (Maziarka et al., 2024), MAT (Maziarka et al., 2020), MolBERT (Fabian et al., 2020), ChemBERTa (Chithrananda et al., 2020), MolFormer (Ross et al., 2022), CDDD (Winter et al., 2019), ECFP (Rogers & Hahn, 2010), Atom Pair (Carhart et al., 1985), Mol2Vec (Jaeger et al., 2018), Uni-Mol2 (Ji et al., 2024), GROVER (Rong et al., 2020), and GraphMVP (Liu et al., 2022). Each is the single best-performing variant of its model family, selected by aggregate AUROC across all 24 datasets from a large-scale fair benchmarking study (Praski et al., 2025), using each model's best classifier head (random forest, logistic regression, or k-NN) per dataset. Full selection details and the complete comparison across 57 model variants are reported in section C.2. We note that CLAMP is pretrained via supervised contrastive learning on ChEMBL bioactivity assay data; its scores on bioactivity-adjacent datasets should be read with this potential domain overlap in mind as discussed in section E.2.

Table 1: **Molecular property prediction results (ROC-AUC).** Comparison of 13 representative baselines from Praski et al. (2025) across Overall, OGB, and TDC benchmarks against MOLFIBER-G, MOLFIBER-L and MOLFIBER-ENS. **Best**, **second** and **third** values are marked.

	Overall (24 datasets)			OGB (6 datasets)			TDC (18 datasets)		
Method	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓
CLAMP	**82.01**	**5.63**	**3.61**	**79.18**	6.42	**5.19**	**82.96**	**5.36**	**3.08**
R-MAT	**80.97**	6.46	4.65	77.75	7.58	6.62	82.04	6.08	3.99
CDDD	80.82	8.17	4.80	78.43	8.00	5.94	81.62	8.22	4.42
MolBERT	80.67	7.63	4.95	76.95	9.67	7.42	81.91	6.94	4.13
MAT	80.38	8.48	5.24	77.60	9.00	6.77	81.31	8.31	4.73
ChemBERTa	80.15	7.92	5.47	75.90	7.83	8.47	81.57	7.94	4.47
ECFP	80.14	7.79	5.48	76.67	6.83	7.70	81.29	8.11	4.74
MolFormer	79.80	9.90	5.83	77.20	9.75	7.17	80.66	9.94	5.38
Mol2Vec	79.59	10.35	6.03	78.12	7.42	6.25	80.08	11.33	5.96
Atom Pair	79.53	8.21	6.09	76.18	**5.08**	8.19	80.64	9.25	5.39
Uni-Mol	78.29	12.08	7.33	73.65	13.58	10.72	79.84	11.58	6.20
GraphMVP	77.65	12.58	7.98	73.82	12.67	10.55	78.92	12.56	7.12
GROVER	76.13	12.94	9.49	70.87	12.00	13.50	77.88	13.25	8.16
MolFiber-G	**82.01**	**5.19**	**3.62**	**80.15**	**6.33**	**4.22**	**82.62**	**4.81**	**3.41**
MolFiber-L	80.96	7.92	4.66	79.10	8.17	5.27	81.58	7.83	4.46
MolFiber-ENS	**82.10**	**4.77**	**3.52**	**80.16**	**5.67**	**4.21**	**82.75**	**4.47**	**3.29**

Implementation details. MolFiber is implemented in PyTorch. The Morgan anchor is a scikit-learn random forest with $N_{\text{trees}} = 200$ on binary ECFP4 fingerprints (radius 2, 2048 bits), scored out of fold via 5-fold cross-fitting on training molecules and refit on the full training pool at inference (**??**). GraphGrid and MolFormer (`ibm/MoLFormer-XL-both-10pct`) are frozen feature extractors; only the projection heads P_G and P_F are trained. We use fiber size $k = 8$ (section D.3), residual bound $\lambda = 4$, latent width $d = 128$ for both the view projections and the pair embedding, and hidden width 256 in φ and ψ, with attention temperatures $\tau = \tau_c = 0.1$. All variants are trained with Adam, learning rate 10^{-3}, weight decay 10^{-5}, batch size 64, for up to 100 epochs. Following the benchmark protocol in Praski et al. (2025), training and validation molecules form a single pool and the epoch is selected by 5-fold cross-validation within it, after which the model is refit on the entire pool; the test split is never used for selection. Results are reported as mean $\pm$ standard deviation over $n = 5$ runs with fixed seeds shared across all methods and baselines. Experiments were run on a single NVIDIA H200 GPU (141 GB). Full hyperparameters, the selection protocol and runtime measurements are given in Appendix C.3. Our code is available at
`https://anonymous.4open.science/r/MolFiber-2F5E`

4.1 MAIN RESULTS

Table 1 reports Mean AUC, Average Rank, and Average Gap across the Overall, OGB, and TDC benchmark collections. MOLFIBER-ENS achieves the strongest Overall performance among all 16 methods evaluated, attaining the highest Overall Mean AUC (82.10), the lowest Overall Average Rank (4.79), and the smallest Overall Average Gap (3.52), indicating that the ensembled variant is both the most accurate and the most consistently competitive across all 24 datasets. On the OGB subset, MOLFIBER-ENS again leads in Mean AUC (80.16) and Average Gap (4.22), while ATOM PAIR attains a marginally lower Average Rank (5.08 vs. 5.83), suggesting that although MOLFIBER-ENS is closer to the per-dataset optimum on average, ATOM PAIR places slightly more consistently across the six individual OGB datasets.

On the TDC subset, CLAMP remains the strongest baseline, achieving the best Mean AUC (82.96) and Average Gap (3.07). Notably, CLAMP does not abandon the fingerprint: it retains a fingerprint-based molecule encoder and adds a second signal, contrastive alignment to textual bioactivity assay descriptions from ChEMBL, rather than replacing it outright. That even the strongest competing method keeps a fingerprint as a load-bearing component is consistent with our central premise that Morgan-based representations remain difficult to improve upon except through careful conditioning of an auxiliary signal. MOLFIBER-ENS nonetheless attains the lowest TDC Average Rank (4.44), and MOLFIBER-A is a close second on both Average Rank (4.72) and Mean AUC (82.67), narrowly trailing CLAMP.

Notably, all three MOLFIBER variants outperform representative pretrained and self-supervised baselines such as UNI-MOL, GRAPHMVP, and GROVER, which exhibit markedly higher Average Gaps (7.33–9.49 Overall), reflecting greater inconsistency across the diverse property-prediction tasks. Taken together, these results indicate the MOLFIBER variants are competitive with or superior to the strongest single-representation baselines. Ensembling MOLFIBER-G and MOLFIBER-L into MOLFIBER-ENS yields the most reliable performance across heterogeneous molecular property prediction benchmarks, closing much of the residual gap to the best per-dataset result. Tables 14 and 15 show the full results for all the datasets.

4.2 ABLATIONS

We use ablations to isolate three questions central to MolFiber: whether its gains come from fiber-relative rather than standard fusion, whether simpler G/L endpoints outperform learned alternatives, and how sensitive performance is to fiber size K.

Why Fiber-Relative Fusion? MolFiber is itself a multiview fusion method, but with a deliberately asymmetric organization: Morgan defines the global geometry, while GraphGrid and MolFormer contribute only through comparisons within Morgan-local neighborhoods. The relevant question is therefore not whether multiple views help, but whether organizing them through the anchor geometry is more effective than standard fusion of the same information.

The controlled comparison in table 2 separates these effects. Individual views are weak in aggregate: GraphGrid reaches 64.95 average AUC, Morgan-RF 72.94, and MolFormer 77.20. Standard fusion substantially improves on these representations, with Early Fusion and Late Fusion reaching 79.22 and 78.98, respectively. Thus, complementary views do contain useful joint information.

MolFiber extracts more from the same views. The single-model MOLFIBER-G reaches 80.15 average AUC, exceeding all three controlled fusion baselines, while MOLFIBER-ENS reaches 80.16 and achieves the best average rank and gap. This is important because the improvement cannot be attributed solely to the two-model ensemble: the single MolFiber-G network already improves over the strongest standard fusion baseline.

The most direct control is Anchor-MLP. It retains the frozen Morgan anchor, the bounded residual formulation, and the same auxiliary representations, but computes its correction from the global embeddings $[g_x, m_x]$ rather than from within-fiber comparisons. Removing the fiber-relative organization drops average AUC from 80.15 for MolFiber-G to 77.68, worsens average rank from 3.33 to 5.17, and increases average gap from 0.50 to 2.97. A strong anchor and bounded correction are therefore not sufficient by themselves; the main gain comes from conditioning auxiliary evidence on Morgan-local structure.

The advantage is not uniform across every dataset. Early Fusion leads HIV and TOX21, while Late Fusion leads SIDER. MolFiber-G/L/ENS nevertheless occupy the top three positions on four of the six datasets and achieve the strongest aggregate AUC, rank, and gap. We report per-dataset uncertainty and statistical comparisons in section D.1.

Table 2: **Individual views, standard fusion, and MolFiber.** AUC on the six OGB datasets. Stars mark MolFiber-ENS significance over individual methods (Fisher's combined test; full statistics in section D.1).

Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	Av.AUC ↑	Av.Rank ↓	Av.Gap ↓
Morgan-RF	85.21	67.96	70.59	79.40	62.95	71.55	72.94***	7.00	7.71
Graph grid	76.24	62.43	63.49	72.64	53.23	61.64	64.95***	9.00	15.70
MolFormer	83.70	**71.50**	88.40	78.00	**66.60**	75.00	77.20*	5.33	3.45
Early Fusion	83.12	71.07	98.29	**80.91**	63.36	**78.56**	**79.22*****	**4.00**	**1.43**
Late Fusion	83.09	67.99	98.30	**80.23**	**66.67**	**77.57**	78.98***	4.17	1.67
Anchor MLP	82.09	70.34	89.89	79.35	**66.41**	**77.99**	77.68***	5.17	2.97
MolFiber-G	**85.45**	**72.53**	**99.17**	80.04	66.25	77.47	**80.15**	**3.33**	**0.50**
MolFiber-L	**85.50**	69.84	**99.31**	79.53	64.90	75.50	79.10***	4.17	1.55
MolFiber-ENS	**85.49**	**72.94**	**99.29**	**80.33**	65.80	77.09	**80.16**	**2.83**	**0.49**

Table 3: **Performance Summary of MolFiber variants.** Average AUC, Rank, and Gap computed among these 6 models. Complete results for individual datasets are given in Table 11.

	Overall			OGB			TDC		
Method	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓
MolFiber-A	81.99	2.88	0.29	79.95	3.50	0.36	82.67	2.67	**0.26**
MolFiber-M	81.50	3.75	0.77	79.99	**2.83**	**0.32**	82.01	4.06	0.92
MolFiber-route	80.87	5.00	1.40	78.62	6.00	1.69	81.63	4.67	1.30
MolFiber-G	**82.01**	**2.44**	**0.27**	**80.15**	**2.50**	**0.16**	**82.62**	**2.42**	0.31
MolFiber-L	80.96	4.79	1.32	79.10	3.67	1.22	81.58	5.17	1.35
MolFiber-ENS	**82.10**	**2.15**	**0.18**	**80.16**	**2.50**	**0.16**	**82.75**	**2.03**	**0.18**

MolFiber variants. Equation 1's family of corrections extends beyond the two endpoints $\rho \in \{0, 1\}$ studied above (Section D.2): **MolFiber-M** learns a single $\rho^{(t)}$ per property, and **MolFiber-A** estimates $\rho^{(t)}(x)$ adaptively for each molecule from local fiber reliability. We additionally compare against **MolFiber-route**, which replaces the joint pair encoder with two separate per-view relative experts over the same fiber $\mathcal{B}_x$ — one for GraphGrid, one for MolFormer — gated by a task-specific router that can defer to either expert or leave the Morgan prediction untouched. None of these three variants exceeds MolFiber-G or MolFiber-ENS in aggregate AUC, rank, or gap across Overall, OGB, or TDC. MolFiber-M is competitive with MolFiber-G specifically on OGB (79.99 vs. 80.15 AUC), but falls off on TDC; MolFiber-A comes closest to MolFiber-ENS overall (82.67 vs. 82.75 TDC AUC) without surpassing it; and MolFiber-route is the weakest variant we study, trailing even MolFiber-L. Learning ρ, whether per task or per molecule, therefore does not improve on fixing it at an endpoint — or better, ensembling both endpoints — on this benchmark suite. Full per-dataset results and aggregate statistics for all six variants are given in Appendix D.2.

Number of neighbors per fiber. We ablate the fiber size K on the OGB datasets (table 4). Most of the gain is achieved by $K = 8$: performance improves from 79.75 at $K = 1$ to 80.16, changes only marginally at $K = 16$, and slightly declines at $K = 24$. We therefore use $K = 8$ throughout: it captures nearly all of the benefit of larger fibers while avoiding the approximately 4–9× higher pairwise computation of $K = 16$ and $K = 24$ under the $O(NK^2)$ scaling discussed in section A.3. Full per-dataset fiber-size results are reported in section D.3, including the per-method breakdown across MolFiber-G, MolFiber-L, and MolFiber-ENS in Table 12.

Table 4: **Fiber size K.** MolFiber-ENS Av. AUC on OGB datasets.

K	Av. AUC ↑
1	79.75
8	**80.16**
16	**80.21**
24	80.08

5 Conclusion

We introduced MolFiber, a fibration-based approach to combining molecular representations in which a frozen Morgan anchor is refined by GraphGrid and MolFormer only through comparisons within approximate Morgan fibers. Decomposing the contextual score into within-fiber and between-fiber components shows that the absolute and neighbourhood-centred corrections are not competing designs but two endpoints of a single family, realised as MolFiber-G and MolFiber-L. Across 24 OGB and TDC benchmarks, MolFiber-G and MolFiber-ENS outperform the Morgan anchor and early and late fusion of the same views on average, and improve markedly over an anchored MLP that keeps the anchor and the bounded residual but removes the fiber, isolating fiber-relative comparison as the source of the gain. MolFiber-G outperforms MolFiber-L, while their ensemble, MolFiber-Ens, is stronger than either, indicating that the two endpoints make complementary errors and that neither alone captures the full correction.

Several directions remain. Different fingerprints disagree substantially about which molecules are neighbours, so routing over multiple fiber geometries, or over the views themselves, is a natural extension. More broadly, the construction requires only a strong anchor that induces meaningful local neighbourhoods and auxiliary views that are informative within them, and should transfer beyond molecular property prediction to settings such as protein function prediction, materials property prediction, or tabular learning, where a trusted base model is refined by weaker or noisier modalities. Extending MolFiber to regression targets and to jointly fine-tuned encoders would further test how far the fibration view generalises.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising portions of the manuscript, primarily to improve English, clarity, organization, and readability; to assist with research-code development and debugging; and to help create and refine schematic pipeline figures. All methodological choices, model designs, experimental protocols, and scientific conclusions were made by the authors. AI-assisted code was reviewed, tested, and verified before use, and all AI-assisted text and visual material was checked and revised by the authors. Generative AI was not used to independently establish the scientific claims or evaluate the validity of the reported results. The authors take full responsibility for the final manuscript, code, experiments, and conclusions.

ETHICS STATEMENT

This work studies methodological questions in molecular property prediction using publicly available molecular benchmark datasets. It does not involve human subjects, private or personally identifiable information, or collection of new biological or clinical data. The experiments use established MoleculeNet and Therapeutics Data Commons benchmarks under their existing research-use protocols.

MolFiber is a general molecular prediction methodology and could in principle be applied in downstream drug-discovery, toxicology, or biomedical settings. Predictions from such models should not be interpreted as experimental or clinical evidence without appropriate domain-specific validation. In particular, practical use in safety-critical molecular design or biomedical decision making would require independent experimental validation and appropriate safeguards. We are not aware of additional ethical concerns specific to the methods or benchmark experiments presented here.

REPRODUCIBILITY STATEMENT

We provide implementation code and experiment scripts in an anonymous repository, including construction of the Morgan anchor, GraphGrid features, MolFormer projections, Morgan–Tanimoto neighborhood retrieval, MolFiber-M, MolFiber-A, MolFiber-ENS, and the global-fusion and routing ablations. The repository also includes training, evaluation, and figure/table generation scripts.

The main paper describes the MolFiber methodology and evaluation protocol, while Appendix A provides the complete architecture, fiber construction, pairwise contextualization, training objective, leakage-control procedure, and computational details. Dataset construction, sizes, task statistics, and benchmark splits are reported in Appendix C.1, and baseline selection and provenance are documented in Appendix C.2. Hyperparameters and experimental settings are selected using held-out validation data, and reported MolFiber results are evaluated over fixed benchmark splits and multiple random seeds.

For the external molecular-representation baselines, we use the fixed model variants and corresponding results from the controlled benchmark study described in Appendix C.2, evaluated on the same dataset pool and splits used here, rather than selecting a different baseline variant for each dataset. Together, the released code, fixed protocols, and appendix details are intended to reproduce the MolFiber experiments, ablations, and reported comparisons. Our code is available at
https://anonymous.4open.science/r/MolFiber-2F5E

REFERENCES

Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. In *ML4Molecules – NeurIPS 2020 Workshop*, 2020.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. In *ML4Molecules – NeurIPS 2020 Workshop*, 2020.

Zachary Fralish, Ashley Chen, Paul Skaluba, and Daniel Reker. DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. *Journal of Cheminformatics*, 15(1): 101, 2023. doi: 10.1186/s13321-023-00769-x.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1263–1272. PMLR, 2017.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 58(1):27–35, 2018.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. Relative molecule self-attention transformer. *Journal of Cheminformatics*, 16(1):3, 2024.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv preprint arXiv:2508.06199*, 2025.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=XrMWUuEevr.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pp. 30458–30490. PMLR, 2023.

Dagmar Stumpfe and Jurgen Bajorath. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012.

Md Joshem Uddin, Soham Changani, Sai Karthik Navuluru, Cuneyt Gurcan Akcora, and Baris Coskunuzer. From nodes to pixels: Topological and structural two-view graph imaging. *NeurIPS*, 2026.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science*, 10(6):1692–1701, 2019.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations (ICLR)*, 2023.

APPENDIX

A MolFiber: Architecture and Training **13**

A.1 Representations, Morgan Anchor, and Fiber Construction 13

A.2 Fiber-Relative Contextualization and MolFiber Predictions 14

A.3 Training and Computational Cost 16

B Alternative MolFiber Variants and Controls **16**

B.1 Learned Shrinkage Variants 16

B.2 Global-Fusion and Routing Controls 17

C Experimental Details **18**

C.1 Datasets 18

C.2 Baselines 19

C.3 Implementation Details 19

C.4 Runtimes 20

D Ablations **21**

D.1 MolFiber vs. Fusion Models 21

D.2 MolFiber Variants: G, L, M, A, ENS, and Route 23

D.3 Number of Neighbors per Fiber (K) 24

E Full Per-Dataset Results **25**

E.1 OGB Results 25

E.2 TDC Results 25

A MOLFIBER: ARCHITECTURE AND TRAINING

A.1 REPRESENTATIONS, MORGAN ANCHOR, AND FIBER CONSTRUCTION

Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

be a molecular dataset with T binary targets, $y_i \in \{0,1\}^T$. Labels may be missing; we write $\mathcal{O}_i \subseteq \{1,\dots,T\}$ for the targets observed for molecule x_i. Each molecule is represented by three complementary views:

$$M(x) \in \{0,1\}^{d_M}, \qquad G(x) \in \mathbb{R}^{C\times\kappa\times\kappa}, \qquad F(x) \in \mathbb{R}^{d_F},$$

corresponding to Morgan fingerprints, GraphGrid, and MolFormer.

Molecular views. For the Morgan view, we use ECFP4 fingerprints (Rogers & Hahn, 2010) with radius 2 and $d_M = 2048$ bits. ECFP summarizes local circular substructures and serves both as the global prediction anchor and as the similarity geometry used to define local fibers.

For the pretrained view, we use the publicly released `MoLFormer-XL-both-10pct` checkpoint (Ross et al., 2022) as a frozen feature extractor. Its molecule-level embedding defines $F(x)$ and is subsequently mapped through a learned projection.

GraphGrid (Uddin et al., 2026) is the graph-structural view used by MolFiber. Let $\mathcal{G}(x) = (V, E)$ be the heavy-atom graph with adjacency matrix A, degree matrix D, and combinatorial Laplacian

$L = D - A$. Each vertex is described by

$$\phi(v) = \left([e^{-tL}]_{vv}, \frac{\deg(v)}{|V| - 1}, \mathrm{core}(v), \mathrm{pr}(v) \right), \tag{5}$$

corresponding to heat-kernel signature, degree centrality, k-core number, and PageRank. These descriptors are quantile-binned within each molecule, and vertices are ordered lexicographically by the resulting bins, with canonical atom ranking used to break ties. If P_σ denotes this permutation, then

$$A^\star = P_\sigma A P_\sigma^\top.$$

The ordered vertices are divided into κ contiguous blocks $B_1, \ldots, B_\kappa$, and the grid records blockwise edge density,

$$G(x)_{ij} = \frac{1}{|B_i||B_j|} \sum_{u \in B_i} \sum_{v \in B_j} A^\star_{uv}. \tag{6}$$

Optional channels may separate edges by bond type. The ordering itself uses graph-organizational descriptors rather than atom identity, making GraphGrid deliberately complementary to the local substructure vocabulary encoded by Morgan fingerprints.

Leakage-free Morgan anchor. For each target t, a random forest $f_M^{(t)}$ is trained on the Morgan fingerprint and produces probability $p_M^{(t)}(x)$. We convert this to the anchor logit

$$b_M^{(t)}(x) = \log \frac{p_M^{(t)}(x)}{1 - p_M^{(t)}(x)}. \tag{7}$$

The Morgan predictor remains frozen throughout MolFiber training.

For training molecules, anchor scores are generated out of fold. We partition the training set into five folds and score each fold using a forest trained on the remaining four. Thus, the label of a training molecule does not enter the Morgan model used to generate its own anchor score. At inference, the forest is refitted on the full training set. This prevents the residual model from being trained against artificially optimistic in-sample Morgan predictions.

Approximate Morgan fibers. Morgan fingerprints also define the local geometry. We use Tanimoto similarity

$$s_M(x, z) = \frac{M(x)^\top M(z)}{\|M(x)\|_1 + \|M(z)\|_1 - M(x)^\top M(z)}$$

and let

$$\mathcal{F}_k(x) = \mathrm{kNN}_M(x; \mathcal{D}_{\mathrm{train}})$$

denote the k most similar training molecules. The corresponding approximate Morgan fiber is

$$\mathcal{B}_x = \{x\} \cup \mathcal{F}_k(x).$$

During training, a molecule is excluded from its own neighbor search; at validation and test time, neighbors are retrieved only from the training set.

A.2 Fiber-Relative Contextualization and MolFiber Predictions

The auxiliary representations are first projected to a common latent width d,

$$g_x = P_G(G(x)) \in \mathbb{R}^d, \qquad m_x = P_F(F(x)) \in \mathbb{R}^d, \tag{8}$$

where P_G is a small convolutional encoder and P_F is a learned projection of the frozen MolFormer embedding. Both projections end in a ReLU.

For every ordered pair $z, w \in \mathcal{B}_x$, we form

$$r_{zw} = \left[g_z - g_w,\ g_z \odot g_w,\ m_z - m_w,\ m_z \odot m_w,\ s_M(z, w) \right] \tag{9}$$

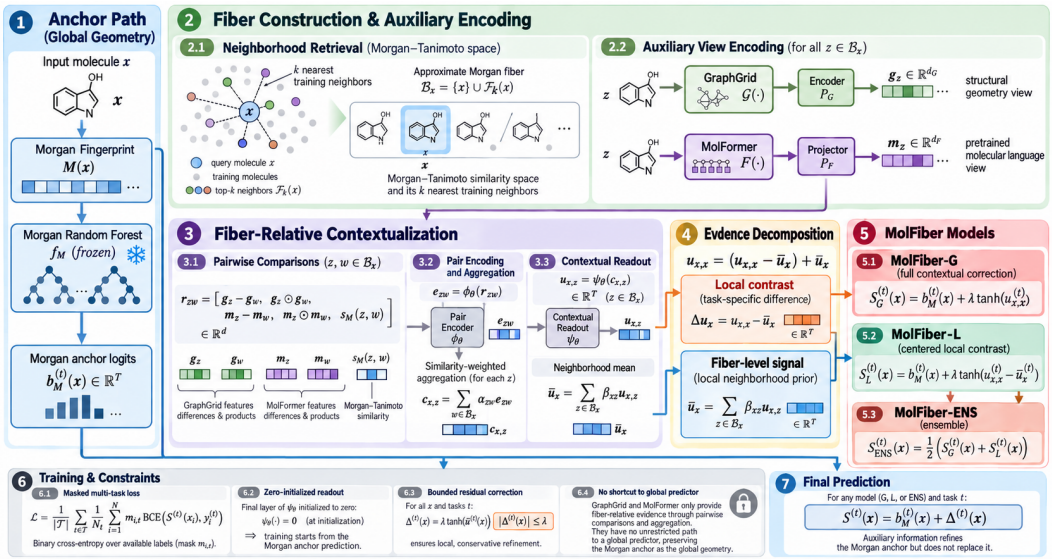

Figure 2: **Detailed MolFiber computational pipeline.** The Morgan path provides a frozen global anchor and defines the Morgan–Tanimoto geometry used to retrieve an approximate fiber around each query molecule. GraphGrid and MolFormer encode complementary structural and pretrained views for all molecules in the fiber. Their pairwise differences, products, and Morgan similarity are processed by a shared pair encoder and similarity-weighted contextual aggregation to produce target-wise auxiliary scores. The query score is then decomposed into a centered local contrast and a fiber-level signal. MolFiber-G retains the full contextual evidence, MolFiber-L uses only the centered local contrast, and MolFiber-ENS averages the independently trained G and L logits. All variants start from the frozen Morgan predictor through zero initialization and add only a bounded residual correction. Training details, leakage control, and computational complexity are given in Appendix A.

and encode it with the shared pair network

$$e_{zw} = \varphi_\theta(r_{zw}) \in \mathbb{R}^{d_e}. \tag{10}$$

For each $z \in \mathcal{B}_x$, pair embeddings are aggregated using Morgan-similarity weights,

$$c_{x,z} = \sum_{w \in \mathcal{B}_x \setminus \{z\}} \alpha_{zw} e_{zw}, \tag{11}$$

$$\alpha_{zw} = \frac{\exp\big(s_M(z,w)/\tau\big)}{\sum_{v \in \mathcal{B}_x \setminus \{z\}} \exp\big(s_M(z,v)/\tau\big)}. \tag{12}$$

A shared contextual readout then produces

$$u_{x,z} = \psi_\theta(c_{x,z}) \in \mathbb{R}^T. \tag{13}$$

Importantly, ψ_θ receives only the fiber-relative context $c_{x,z}$; it does not receive the absolute GraphGrid embedding g_x, MolFormer embedding m_x, or Morgan anchor $b_M(x)$. Thus the auxiliary views have no unrestricted path to a competing global property predictor.

We summarize the neighborhood evidence by

$$\bar{u}_x = \sum_{z \in \mathcal{F}_k(x)} \beta_{xz} u_{x,z}, \tag{14}$$

$$\beta_{xz} = \frac{\exp\big(s_M(x,z)/\tau_c\big)}{\sum_{v \in \mathcal{F}_k(x)} \exp\big(s_M(x,v)/\tau_c\big)}. \tag{15}$$

The query score therefore admits the decomposition

$$u_{x,x} = \underbrace{(u_{x,x} - \bar{u}_x)}_{\text{local contrast}} + \underbrace{\bar{u}_x}_{\text{fiber-level signal}}. \tag{16}$$

This yields our two complementary MolFiber predictors. MolFiber-G retains the full fiber-conditioned auxiliary score,

$$S_G^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\big(u_{x,x}^{(t)}\big), \tag{17}$$

whereas MolFiber-L removes the fiber-level component and uses only the within-fiber contrast,

$$S_L^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\big(u_{x,x}^{(t)} - \bar{u}_x^{(t)}\big). \tag{18}$$

Thus MolFiber-G preserves both local contrast and signal shared by the local Morgan neighborhood, while MolFiber-L uses a fully relative correction that is invariant to a common shift in contextual scores within the fiber.

Our strongest model, MolFiber-ENS, independently trains MolFiber-G and MolFiber-L and averages their prediction logits,

$$S_{\mathrm{ENS}}^{(t)}(x) = \frac{1}{2}\left(S_G^{(t)}(x) + S_L^{(t)}(x)\right). \tag{19}$$

The ensemble combines the complementary absolute and relative uses of fiber-conditioned evidence without introducing an additional gating model.

A.3 Training and Computational Cost

All MolFiber variants are trained with masked binary cross-entropy,

$$\mathcal{L} = -\frac{1}{\sum_i |\mathcal{O}_i|} \sum_i \sum_{t \in \mathcal{O}_i} \Big[y_{it} \log \sigma\big(S^{(t)}(x_i)\big) + (1 - y_{it}) \log\big(1 - \sigma(S^{(t)}(x_i))\big)\Big], \tag{20}$$

where missing targets contribute no loss.

The Morgan predictor is frozen. The trainable components are the GraphGrid encoder, MolFormer projection, pair encoder, and contextual readout. The final layer of ψ_θ is initialized to zero, giving

$$u_{x,z} = 0 \qquad \Longrightarrow \qquad S^{(t)}(x) = b_M^{(t)}(x)$$

at initialization. Thus every model starts exactly from the Morgan anchor and learns only residual corrections supported by the auxiliary views. Furthermore, the $\tanh$ parameterization guarantees

$$\left|S_G^{(t)}(x) - b_M^{(t)}(x)\right| \le \lambda, \qquad \left|S_L^{(t)}(x) - b_M^{(t)}(x)\right| \le \lambda.$$

The same bound also holds for their averaged ensemble.

GraphGrid and MolFormer representations are computed once per molecule. For a fiber containing $k+1$ molecules, MolFiber evaluates ordered pairwise interactions among its members, requiring $O(Nk^2)$ pair-encoder evaluations for N molecules. For fixed small k, the neural computation is therefore linear in dataset size. MolFiber-ENS independently trains and evaluates the G and L networks and consequently requires approximately twice the network cost of a single endpoint model.

B Alternative MolFiber Variants and Controls

The main paper focuses on MolFiber-G, MolFiber-L, and MolFiber-ENS. Here we describe additional variants that test whether the relationship between the two endpoint corrections should instead be learned, adapted per molecule, or reorganized architecturally.

B.1 Learned Shrinkage Variants

The decomposition in eq. (16) defines a continuous family between MolFiber-L and MolFiber-G:

$$S_\rho^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\left[\big(u_{x,x}^{(t)} - \bar{u}_x^{(t)}\big) + \rho\, \bar{u}_x^{(t)}\right]. \tag{21}$$

Equivalently,

$$S_\rho^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\left[u_{x,x}^{(t)} - (1-\rho)\bar{u}_x^{(t)}\right].$$

The endpoints recover MolFiber-L at $\rho = 0$ and MolFiber-G at $\rho = 1$.

MolFiber-M. MolFiber-M learns one retention coefficient per target,

$$\rho^{(t)} = \sigma\big(\varrho^{(t)}\big), \qquad \varrho \in \mathbb{R}^T. \tag{22}$$

The coefficient is shared across molecules but can differ across properties, allowing each task to determine how much fiber-level signal should be retained.

MolFiber-A. MolFiber-A instead makes retention molecule dependent. For target t, we measure within-fiber disagreement by

$$s_t^2(x) = \sum_{z \in \mathcal{F}_k(x)} \beta_{xz} \left(u_{x,z}^{(t)} - \bar{u}_x^{(t)} \right)^2, \tag{23}$$

and define the entropy-based effective neighborhood size

$$n_{\text{eff}}(x) = \exp\left(- \sum_{z \in \mathcal{F}_k(x)} \beta_{xz} \log \beta_{xz} \right). \tag{24}$$

The adaptive retention coefficient is

$$\rho^{(t)}(x) = \frac{\tau_t^2}{\tau_t^2 + s_t^2(x)/n_{\text{eff}}(x)}, \tag{25}$$

where

$$\tau_t^2 = \kappa_t \, \widehat{\text{Var}}_x \left(\bar{u}_x^{(t)} \right), \qquad \kappa_t > 0. \tag{26}$$

The variance is maintained as a running estimate during training and frozen at inference. Hence coherent fibers yield $\rho^{(t)}(x)$ near one, while fibers with large local disagreement move toward the centered MolFiber-L correction.

These learned variants test whether explicitly estimating intermediate retention between the G and L endpoints improves over using the two simpler inductive biases directly. Their empirical comparison is reported in table 1 and the corresponding appendix tables.

B.2 Global-Fusion and Routing Controls

Anchor-MLP global control. A central question is whether MolFiber benefits specifically from organizing auxiliary information within Morgan-local neighborhoods, rather than simply from combining the same representations. We therefore construct a global control that retains the frozen Morgan anchor and the same projected GraphGrid and MolFormer representations, but removes the fiber construction.

The auxiliary embeddings are combined directly,

$$h_x = h_\theta([g_x, m_x]), \tag{27}$$

and used to form the bounded correction

$$S_{\text{GF}}^{(t)}(x) = b_M^{(t)}(x) + \lambda \tanh\big(h_x^{(t)}\big). \tag{28}$$

This control preserves the anchor and access to both auxiliary views while removing Morgan-local contextualization. Comparing it with MolFiber therefore isolates the contribution of the fiber-relative organization itself.

Relative-expert router. We also test whether GraphGrid and MolFormer benefit from acting as separate local experts rather than contributing jointly to the same pair encoder. The GraphGrid expert receives

$$[g_z - g_w,\ g_z \odot g_w,\ s_M(z,w)],$$

while the MolFormer expert receives

$$[m_z - m_w,\ m_z \odot m_w,\ s_M(z,w)].$$

Each produces a Morgan-local relative score,

$$R_G^{(t)}(x), \qquad R_F^{(t)}(x).$$

A task-specific router assigns probabilities to the three alternatives

$$\{\text{null}, \text{GraphGrid}, \text{MolFormer}\},$$

with

$$\left(\pi_0^{(t)}(x), \pi_G^{(t)}(x), \pi_F^{(t)}(x)\right) = \text{softmax}\left(q^{(t)}(x)\right). \tag{29}$$

The resulting correction is

$$\Delta_R^{(t)}(x) = \lambda \left[\pi_G^{(t)}(x) \tanh R_G^{(t)}(x) + \pi_F^{(t)}(x) \tanh R_F^{(t)}(x)\right], \tag{30}$$

while the null route contributes zero.

The router therefore tests whether explicitly specializing the two auxiliary representations as separate local experts improves over their joint fiber-relative encoding. We treat it as an architectural ablation rather than a primary MolFiber model.

Table 5: **Datasets.** Number of molecules, tasks, and mean share of positive labels across tasks, following the dataset construction of Praski et al. (2025).

Dataset	Source	# Molecules	# Tasks	% pos.
BACE	MoleculeNet	1,513	1	45.67%
BBBP	MoleculeNet	2,039	1	76.51%
ClinTox	MoleculeNet	1,477	2	50.61%
HIV	MoleculeNet	41,084	1	3.50%
SIDER	MoleculeNet	1,427	27	56.76%
Tox21	MoleculeNet	7,831	12	7.52%
AMES	TDC	7,278	1	54.60%
Bioavailability_Ma	TDC	640	1	76.88%
CYP1A2_Veith	TDC	12,579	1	46.34%
CYP2C19_Veith	TDC	12,665	1	45.95%
CYP2C9_Substrate_CarbonMangels	TDC	669	1	21.08%
CYP2C9_Veith	TDC	12,092	1	33.45%
CYP2D6_Substrate_CarbonMangels	TDC	667	1	28.64%
CYP2D6_Veith	TDC	13,130	1	19.15%
CYP3A4_Substrate_CarbonMangels	TDC	670	1	52.99%
CYP3A4_Veith	TDC	12,328	1	41.45%
DILI	TDC	475	1	49.68%
HIA_Hou	TDC	578	1	86.51%
PAMPA_NCATS	TDC	2,034	1	85.50%
Pgp_Broccatelli	TDC	1,218	1	53.37%
SARSCoV2_3CLPro_Diamond	TDC	880	1	8.86%
SARSCoV2_Vitro_Touret	TDC	1,484	1	5.93%
hERG	TDC	655	1	68.85%
hERG_Karim	TDC	13,445	1	49.97%

C Experimental Details

C.1 Datasets

Table 5 lists all 24 benchmarks used in this work: 6 classification datasets from MoleculeNet (Wu et al., 2018), evaluated under the Open Graph Benchmark scaffold splits (Hu et al., 2020), and 18 ADME/toxicity endpoints from Therapeutics Data Commons (Huang et al., 2021), evaluated under TDC's own scaffold-split protocol. For consistency with those baseline scores, we exclude the same small set of molecules excluded in the source benchmark – fewer than 0.01% of MoleculeNet HIV due to SELFIES-translation parsing errors in some of their baseline models, and fewer than 0.003% containing dative bonds that a subset of their baseline models could not process – listed in full in their released repository.

Three datasets are multitask: ClinTox (2 tasks), Tox21 (12), and SIDER (27); the remaining 21 are single-task. We use the masked multitask formulation described in section A.3 throughout, training a single model instance per dataset rather than per task.

Table 6: **Baseline models.** For each of 13 architectural categories, we select the single best-performing variant (by mean AUROC across the 24 evaluation datasets) as a fixed baseline, following the protocol and raw per-variant results of Praski et al. (2025); the remaining variants within each model family are reported there.

Category	Model	Best variant	Mean AUC	Why include
Classical fingerprint	**ECFP** (Rogers & Hahn, 2010)	`ECFP_count`	0.8014	Mandatory reference point.
Classical fingerprint (2nd)	**Atom Pair** (Carhart et al., 1985)	`AtomPair_count`	0.7953	Shows the fingerprint family isn't a one-off.
Fingerprint + Word2Vec hybrid	**Mol2Vec** (Jaeger et al., 2018)	`mol2vec`	0.7960	Distinct hybrid paradigm.
Fingerprint fusion + contrastive	**CLAMP** (Seidl et al., 2023)	`CLAMP`	0.8201	Strongest overall in this benchmark.
Graph transformer	**R-MAT** (Maziarka et al., 2024)	`rmat_4M`	0.8098	2nd strongest; relative-attention successor to MAT.
Graph transformer (predecessor)	**MAT** (Maziarka et al., 2020)	`mat_masking_2M`	0.8038	Distance-aware attention, distinct lineage from R-MAT, well cited.
SMILES transformer (dual-obj.)	**MolBERT** (Fabian et al., 2020)	`molbert`	0.8067	MLM + property regression.
SMILES transformer (RoBERTa)	**ChemBERTa** (Chithrananda et al., 2020)	`ChemBERTa-10M-MTR`	0.8015	Most recognizable SMILES-LM baseline.
SMILES transformer (linear attn.)	**MolFormer** (Ross et al., 2022)	`MoLFormer-XL-both-10pct`	0.7980	One of MolFiber's own two auxiliary views
SMILES RNN autoencoder	**CDDD** (Winter et al., 2019)	`CDDD`	0.8083	Earlier-generation text approach, translation objective
3D conformer transformer	**Uni-Mol2** (Ji et al., 2024)	`unimolv2`	0.7829	Near-mandatory comparison point despite lower score here.
Hybrid GNN-transformer	**GROVER** (Rong et al., 2020)	`grover_large`	0.7613	Near-mandatory comparison point despite lower score here.
GNN self-sup. pretraining	**GraphMVP** (Liu et al., 2022)	`GraphMVP_CP-max`	0.7766	Represents 2D/3D-contrastive GNN pretraining

C.2 Baselines

Source and selection protocol. All 13 baselines in our experiments are drawn from Praski et al. (2025)'s large-scale, fair comparison of pretrained and classical molecular representation models. Their study evaluates **57 embedder variants** under all three classifier heads (random forest, logistic regression, and k-NN) on the same 24 datasets used in this work; most of these 57 variants are different sizes, pretraining objectives, or architectural configurations of the same 13 underlying model families that constitute our baselines.

Reporting a per-dataset maximum over all 57 variants would import multiple-comparison inflation into a single "best baseline" row, effectively comparing our models against a different oracle-selected winner on every column. We avoid this by fixing one representative variant per model family, selected by its aggregate mean AUROC across all 24 datasets, and using that single variant's score, with, for each dataset, the best-performing classifier head, following Praski et al. (2025)'s reported protocol, consistently across every column. Each baseline in table 6 is therefore a fixed, reproducible model, not a shifting oracle.

Category coverage. Table 6 lists the 13 selected baselines by architectural category, with the rationale for including each. Two entries, Uni-Mol2 (Ji et al., 2024) and GROVER (Rong et al., 2020), are included despite lower aggregate AUROC than several excluded alternatives: both are near-mandatory comparison points in molecular representation learning, and reporting their true, lower scores here is itself consistent with Praski et al. (2025)'s central finding that widely-assumed state-of-the-art performance does not always reproduce under fair comparison.

Full variant set. The complete set of 57 embedder variants – including all sizes and pretraining objectives excluded from Table 6 for space (e.g. ChemBERTa at 5M/10M/77M parameters under both MLM and MTR objectives, MAT at 200k/2M/20M parameters, GraphMVP's six SSL/pooling combinations, and both Uni-Mol (Zhou et al., 2023) and Uni-Mol2 (Ji et al., 2024)) – along with their raw per-dataset, per-classifier-head results, is reported in full in Praski et al. (2025).

C.3 Implementation Details

All experiments use the default configuration of the released code unless stated otherwise; no per-dataset tuning was performed. Code is available at `https://anonymous.4open.science/r/MolFiber-2F5E`.

Views. The Morgan view is a binary ECFP4 fingerprint (radius 2, 2048 bits, RDKit `MorganGenerator`), used both to train the anchor and to retrieve fibers. GraphGrid uses a 10×10 grid with a single adjacency channel: per molecule we compute the heat kernel signature at $t = 0.1$, degree centrality, k-core number and PageRank ($\alpha = 0.85$) on the unweighted heavy-atom graph, quantile-bin each descriptor into 10 bins within the molecule, order vertices lexicographically with ties broken by vertex index, and block-pool the reordered adjacency into equal contiguous blocks. Descriptors are computed in float64 and rounded to 10^{-10} before binning: the binning makes the ordering sensitive to numerical noise, and a float32 eigendecomposition can otherwise move a vertex across a bin edge and change the entire image. Grid channels are standardised using training-split statistics only. The MolFormer view is the mean-pooled final hidden state of `ibm/MoLFormer-XL-both-10pct` over non-padding tokens (768-dimensional), computed once and cached; the checkpoint is frozen throughout.

Anchor. A random forest (200 trees, unlimited depth, minimum leaf size 1) is fitted per target. On training molecules b_M is obtained out of fold by 5-fold cross-fitting, so no molecule's own label contributes to its anchor score; at inference the forest is refit on the entire training pool. Targets with a single observed class are skipped. The anchor is frozen and receives no gradient.

Fibers. For every molecule, train or test, the $k = 8$ nearest neighbours are retrieved from the training pool under Morgan–Tanimoto similarity, giving $|\mathcal{B}_x| = k + 1 = 9$. Retrieval is always from training data, so the method is inductive and no test molecule is compared against another test molecule; training molecules are excluded from their own neighbour sets. Retrieval is label-free.

Model and optimisation. P_G is a two-layer CNN (widths 32 and 64, 3×3 kernels, batch normalisation, mean- and max-pooling) and P_F a linear projection, both with ReLU outputs and mapping to $d = 128$. The pair encoder φ takes the $4d + 1 = 513$-dimensional pair descriptor through a hidden layer of width 256 to a 128-dimensional pair embedding; the readout ψ maps it through width 256 to the per-target score. Dropout is 0.1. Attention temperatures are $\tau = \tau_c = 0.1$ and the residual bound is $\lambda = 4$, held fixed. The output layer of ψ is zero-initialised, so $S(x) = b_M(x)$ identically at initialisation for every variant. Training uses Adam (learning rate 10^{-3}, weight decay 10^{-5}), batch size 64, gradient-norm clipping at 5, and masked binary cross-entropy in which unobserved targets contribute nothing.

Selection protocol. We follow the benchmark's protocol given in Praski et al. (2025), which provides a training pool and a held-out test set and no separate validation fold. Model selection therefore uses 5-fold cross-validation inside the training pool: each fold is trained on the remaining four with early stopping (patience 15, maximum 100 epochs) on the held-out fold, and the median best epoch across folds is used to refit a final model on the entire pool with early stopping disabled. This is the neural analogue of `refit=True` in the grid search used by the benchmark's reference implementation. The test split is used exactly once, to score the refit model. Reported numbers are the mean and standard deviation over $n = 5$ seeds, which vary model initialisation and batch order; the anchor and all cached features are computed once and shared across seeds and across all methods, so every method sees byte-identical splits, fingerprints, anchor logits, GraphGrid tensors and MolFormer embeddings.

Hardware and runtime. All experiments were run on a single NVIDIA H200 GPU with 141 GB of memory. Per-molecule cost is quadratic in fiber size, since the pair encoder evaluates all $(k + 1)^2$ ordered pairs within a fiber; at $k = 8$ this is 81 pairs per molecule per step. Feature construction is a one-off cost per dataset (fingerprints, GraphGrid images, MolFormer embeddings and within-fiber similarity blocks are all cached and reused across variants, seeds and baselines), leaving the pair encoder as the dominant per-epoch term.

C.4 Runtimes

Table 7 reports wall-clock runtime in seconds for each stage of the MolFiber pipeline on three OGB datasets spanning nearly two orders of magnitude in size ($n = 1{,}513$ to $41{,}120$). The first four columns (fiber construction, GraphGrid, MolFormer, and Morgan-RF) are upstream costs computed once per dataset and shared across all MolFiber variants: GraphGrid and MolFormer are

Table 7: Runtime analysis in seconds for three OGB datasets.

Dataset	n	Tasks	Fiber Construction	GraphGrid	MolFormer	Morgan-RF	MolFiber-G	MolFiber-L	MolFiber-ENS
BACE	1,513	1	0.13	1.37	134.72	3.95	10.96	10.81	21.78
TOX21	7,823	12	2.59	4.78	775.61	369.11	56.45	56.48	112.93
HIV	41,120	1	133.91	27.77	3930.02	607.66	297.61	297.94	595.56

frozen encoders, Morgan-RF is the frozen anchor, and fiber construction depends only on Morgan fingerprints, so none of this is repeated when training MolFiber-G, MolFiber-L, or MolFiber-ENS. Among these shared stages, MolFormer dominates throughout, accounting for 67–96% of upstream cost across the three datasets (e.g. 3930.02s of HIV's 4699.36s upstream total). Fiber construction scales with dataset size as expected from the $O(Nk^2)$ pairwise cost discussed in section A.3 — rising from 0.13s on BACE to 133.91s on HIV, a roughly $1{,}030\times$ increase for a $27\times$ increase in n — but remains a small fraction of the upstream total even at the largest scale tested. Given these precomputed inputs, MolFiber-G and MolFiber-L cost essentially the same to train and evaluate at every scale (e.g. 10.96s vs. 10.81s on BACE, 297.61s vs. 297.94s on HIV), consistent with the two variants sharing an identical architecture and differing only in whether the fiber-level mean is subtracted before the $\tanh$ nonlinearity (eqs. (2) and (3)). MolFiber-ENS's cost is, to within rounding, exactly the sum of G's and L's (e.g. 595.56s vs. 297.61s + 297.94s = 595.55s on HIV), since it trains both independently rather than sharing computation between them — but a practitioner who wants both G and L individually, not just their ensemble, pays this combined cost regardless of whether ENS is used, since the upstream representations are shared either way.

D Ablations

D.1 MolFiber vs. Fusion Models

MolFiber is itself a multiview fusion method. Its distinguishing feature is not access to additional representations, but how those representations are organized: Morgan defines the reference geometry, while GraphGrid and MolFormer contribute through comparisons within Morgan-local fibers. The controlled fusion experiments in table 2 therefore test a central question of the paper: whether this fiber-relative organization improves over standard ways of combining the same molecular views.

The aggregate comparison favors MolFiber. Among the controlled fusion baselines, Early Fusion is strongest at 79.22 average AUC, compared with 80.15 for the single-model MolFiber-G and 80.16 for MolFiber-ENS. The MolFiber-G result is particularly important because it shows that the gain cannot be explained simply by ensembling two models. MolFiber-ENS provides the strongest final result, but a single fiber-relative model already outperforms the standard fusion alternatives in average AUC.

The most direct architectural control is Anchor-MLP. It retains the frozen Morgan anchor, the same GraphGrid and MolFormer representations, and the same bounded-residual formulation, but computes its correction from the global auxiliary embeddings rather than from within-fiber comparisons. Its lower aggregate performance therefore isolates the contribution of the fiber construction more directly than generic early- or late-fusion comparisons.

Test setup. Because aggregate AUC alone does not account for run-to-run variability, we additionally report statistical comparisons using the five independent runs in table 8. For each OGB dataset and each competitor, we compare the MolFiber-ENS mean and standard deviation with the corresponding competitor using Welch's t-test, which does not assume equal variance.

The experiments use the same dataset splits and seed set across methods. However, individual seed-level values were not retained for all models, so a paired test cannot be reconstructed. We therefore compute Welch's test from the reported means, standard deviations, and $n = 5$ runs per method and treat the resulting analysis as an unpaired comparison based on the available summary statistics.

For each competitor, we apply Holm–Bonferroni correction across the six per-dataset tests at $\alpha = 0.05$. We additionally report Fisher's combined p-value as an omnibus summary of whether the two methods differ across the six datasets. Because Fisher's method combines two-sided p-values without using

Table 8: **Individual views, fusion baselines, and MolFiber.** AUC (%) results, mean ± standard deviation over $n = 5$ runs, on OGB datasets.

Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	Average AUC
Morgan-RF	85.21±0.0	67.96±0.0	70.59±0.0	79.40±0.0	62.95±0.0	71.55±0.0	72.94
Graph grid	76.24±1.5	62.43±1.1	63.49±3.1	72.64±2.2	53.23±1.7	61.64±0.9	64.94
MolFormer	83.70±6.8	71.50±2.7	88.40±15.9	78.00±1.3	66.60±1.7	75.00±1.8	77.20
Early Fusion	83.12±1.3	71.07±0.7	98.29±0.4	80.91±0.5	63.36±1.2	78.56±0.2	79.22
Late Fusion	83.09±1.7	67.99±0.7	98.30±0.2	80.23±0.3	66.67±0.3	77.57±0.3	78.98
Anchor MLP	82.09±2.7	70.34±0.6	89.89±0.3	79.35±0.2	66.41±0.9	77.99±0.3	77.68
MolFiber-G	85.45±0.0	72.53±0.7	99.17±0.3	80.04±0.1	66.25±0.3	77.47±0.3	80.15
MolFiber-L	85.50±0.2	69.84±1.2	99.31±0.1	79.53±0.4	64.90±0.2	75.50±0.2	79.10
MolFiber-ENS	85.49±0.1	72.94±0.7	99.29±0.1	80.33±0.5	65.80±0.2	77.09±0.3	80.16

the direction of the AUC difference, this quantity should be interpreted as evidence of an aggregate *difference*, not by itself as a directional test that MolFiber-ENS is superior.

Table 9: **Statistical significance of MolFiber-ENS vs. each baseline.** Per-dataset p-values from Welch's t-test ($n = 5$ runs per method); **bold** indicates Molfiber-ENS significance over the specified method at $\alpha = 0.05$ after Holm–Bonferroni correction across the six datasets. Underlined values indicate significance of the method over MolFiber-ENS. The final column combines all six per-dataset tests via Fisher's method into a single aggregate significance level for that competitor.

Competitor	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	Combined (Fisher)
Morgan-RF	**.003**	**9.1e-5**	**3.5e-11**	**.014**	**5.8e-6**	**2.1e-6**	**<1e-15**
GraphGrid	**<1e-3**	**5.6e-7**	**1.3e-5**	**.001**	**6.7e-5**	**3.9e-7**	**<1e-15**
MolFormer	.588	.306	.200	.013	.354	.059	**.027**
Early Fusion	**.015**	**.003**	**.004**	.104	.010	4.0e-5	**2.5e-9**
Late Fusion	.034	**3.7e-6**	**6.9e-5**	.714	.001	.035	**1.3e-10**
Anchor MLP	.048	**<1e-3**	**2.1e-8**	**.009**	.207	.001	**8.2e-13**

Interpretation. The dataset-level tests reinforce the main aggregate result while also showing that MolFiber's advantage is not universal. Against Early Fusion, MolFiber-ENS is significantly better after Holm correction on BACE, BBBP, ClinTox, and SIDER, while Early Fusion is significantly better on TOX21 and the difference on HIV is not significant. Against Late Fusion, MolFiber-ENS has significant advantages on BBBP and ClinTox, whereas Late Fusion has a significant advantage on SIDER. Against Anchor-MLP, MolFiber-ENS is significantly better on BBBP, ClinTox, and HIV, whereas Anchor-MLP is significantly better on TOX21. Thus, the controlled comparisons contain genuine dataset-specific reversals rather than uniform dominance. Overall, MolFiber-ENS has significant advantages for all the methods when combined over all the datasets.

The Fisher combined tests are highly significant for all three controlled fusion baselines, confirming that the differences between MolFiber-ENS and standard fusion are not well explained by run-to-run variability alone. Directionally, the aggregate performance metrics favor MolFiber-ENS: it achieves higher average AUC than Early Fusion, Late Fusion, and Anchor-MLP, and also the best average rank and average gap in table 2. Together with the single-model MolFiber-G result and the targeted Anchor-MLP control, these comparisons support the conclusion that organizing auxiliary representations through Morgan-local fibers provides useful information beyond conventional global fusion.

The MolFormer comparison is noisier. MolFormer exhibits substantially larger run-to-run variability on several datasets, particularly BACE and ClinTox, so even large differences in mean AUC can have weak dataset-level statistical support. This illustrates why effect size and uncertainty should be considered separately. Our central comparison is therefore with the controlled fusion baselines, which use the same underlying molecular views and directly test MolFiber's fiber-relative organization.

Table 10: **MolFiber variants across Overall, OGB, and TDC.** Av.Rank and Av.Gap computed among these 6 variants.

	Overall			OGB			TDC		
Method	Av.AUC ↑	Av.Rank ↓	Av.Gap ↓	Av.AUC ↑	Av.Rank ↓	Av.Gap ↓	Av.AUC ↑	Av.Rank ↓	Av.Gap ↓
MolFiber-M	81.50	3.75	0.77	79.99	2.83	0.32	82.01	4.06	0.92
MolFiber-A	81.99	2.88	0.29	79.95	3.50	0.36	82.67	2.67	0.26
MolFiber-route	80.87	5.00	1.40	78.62	6.00	1.69	81.63	4.67	1.31
MolFiber-G	82.01	2.44	0.27	80.15	2.50	0.16	82.62	2.42	0.31
MolFiber-L	80.96	4.79	1.32	79.10	3.67	1.22	81.58	5.17	1.35
MolFiber-ENS	82.10	2.15	0.18	80.16	2.50	0.16	82.75	2.03	0.18

Table 11: Per-dataset ROC-AUC for MolFiber variants.

Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	AMES	Bioav.	CYP1A2	CYP2C19	CYP2C9-Sub	CYP2C9
MolFiber-G	85.45	72.53	99.17	80.04	66.25	77.47	84.89	71.27	94.00	89.18	64.53	88.94
MolFiber-L	85.50	69.84	99.31	79.53	64.90	75.50	84.32	68.67	93.63	87.99	62.71	87.66
MolFiber-M	85.47	72.25	99.39	80.00	65.94	76.90	84.56	69.56	93.90	89.00	65.19	88.96
MolFiber-A	85.43	72.95	98.95	79.30	65.92	77.17	84.81	71.03	93.99	89.11	65.66	88.95
MolFiber-ENS	85.49	72.94	99.29	80.33	65.80	77.09	84.87	71.34	94.15	89.16	64.72	89.02
MolFiber-route	85.00	69.83	98.64	78.82	64.50	74.94	84.28	67.78	93.49	88.09	62.76	87.77

Method	CYP2D6-Sub	CYP2D6	CYP3A4-Sub	CYP3A4	DILI	HIA	PAMPA	Pgp	SARS-3CL	SARS-Vitro	hERG	hERG-K
MolFiber-G	80.93	88.47	64.44	89.08	93.76	98.17	80.50	93.18	74.63	63.40	78.84	89.02
MolFiber-L	78.27	87.74	63.93	87.30	92.53	96.59	80.68	92.18	74.36	62.59	79.39	87.92
MolFiber-M	79.78	88.08	64.12	88.94	93.03	98.03	80.49	92.91	74.39	57.78	78.73	88.67
MolFiber-A	80.93	88.42	64.17	88.98	93.56	98.87	80.25	93.34	74.63	63.48	79.14	88.70
MolFiber-ENS	80.65	88.52	64.72	88.82	93.58	98.11	80.86	93.20	74.41	65.05	79.22	89.02
MolFiber-route	78.43	87.59	64.00	87.31	92.32	96.60	81.00	92.05	74.82	63.99	78.92	88.05

D.2 MolFiber Variants: G, L, M, A, ENS, and Route

This ablation studies six points in the shrinkage family $S_\rho(x) = b_M(x) + \lambda \tanh\big(u_{x,x} - (1-\rho)\bar{u}_x\big)$ introduced in section 3: the two fixed endpoints MolFiber-G ($\rho = 1$, retaining the full auxiliary score) and MolFiber-L ($\rho = 0$, retaining only the within-fiber contrast); the two learned variants MolFiber-M (a single $\rho^{(t)}$ per task) and MolFiber-A ($\rho^{(t)}(x)$ estimated per molecule via the empirical-Bayes reliability ratio); the simple ensemble MolFiber-ENS, which averages G's and L's logits; and MolFiber-route, the per-view relative-expert routing ablation described in section B.2. Aggregate results across Overall, OGB, and TDC appear in table 10, with the full 24-dataset breakdown in table 11.

The two endpoints. MolFiber-G outperforms MolFiber-L substantially and consistently: 82.01 vs. 80.96 Overall AUC, 2.44 vs. 4.79 average rank, and a 5× smaller average gap (0.27 vs. 1.32). The gap between the endpoints is narrower on OGB (ranks 2.50 vs. 3.67) than on TDC (2.42 vs. 5.17), where MolFiber-L falls to the second-worst average rank of all six variants. Since G retains the full fiber-level signal and L discards it entirely, this indicates that, in aggregate, fiber-level information is more often useful than harmful on this benchmark suite, discarding it, as L does unconditionally, loses more than it protects against miscalibration.

Learned Shrinkage: Molfiber-M and Molfiber-A. Neither learned variant exceeds MolFiber-G in aggregate. MolFiber-M's OGB rank (2.83) sits close to G's (2.50), suggesting a single task-level ρ recovers most of G's benefit on OGB specifically, but its TDC rank (4.06) is substantially worse than G's (2.42) or A's (2.67); this is driven in part by individual datasets where the learned task-level ρ underperforms sharply, e.g. SARS-Vitro, where MolFiber-M reaches only 57.78 AUC against 62.6–65.1 for every other variant. MolFiber-A is more balanced across the two benchmark subsets (Overall rank 2.88, close to G's 2.44) and is in fact the strongest non-ensemble variant on TDC by raw AUC (82.67) and gap (0.26), narrowly ahead of G. Even so, neither M's per-task nor A's per-molecule estimate of ρ improves on simply fixing $\rho = 1$ in aggregate, the additional estimation machinery does not pay for itself relative to the fixed endpoint on this benchmark suite.

Molfiber-route. MolFiber-route is the weakest variant in every block, trailing even MolFiber-L on Overall and OGB Av.Rank and Av.Gap. Splitting GraphGrid and MolFormer into separate relative experts with a null/GraphGrid/MolFormer router therefore performs worse than combining them

jointly in a single pair encoder (as G, L, M, A, and ENS all do), indicating that the two auxiliary views are better exploited together than routed between.

Ensemble. Despite MolFiber-L's individually weak performance, MolFiber-ENS, the plain average of G's and L's logits — achieves the best Av.AUC, Av.Rank, and Av.Gap of all six variants in every block (Overall, OGB, and TDC). This is not simply G's performance carrying the average: ENS improves on G alone in all nine summary columns, including a lower average gap (0.18 vs. 0.27 Overall) and rank (2.15 vs. 2.44). The two endpoints evidently make different, partially uncorrelated errors, G occasionally over-retains fiber-level signal, L discards it even when it would help, and averaging their predictions cancels some of this error even though neither endpoint alone is closer to optimal than the other's complement.

Taken together, this ablation motivates reporting MolFiber-G, MolFiber-L, and MolFiber-ENS as the primary points of the shrinkage family in the main text, G and L as its two interpretable, fixed endpoints, and ENS as the strongest empirical combination of them, while MolFiber-M, MolFiber-A, and MolFiber-route are reported here as additional variants that do not exceed this simpler combination in aggregate on our benchmark suite.

D.3 Number of Neighbors per Fiber (K)

MolFiber constructs each molecule's fiber $\mathcal{F}_K(x)$ from its K nearest neighbors under Morgan–Tanimoto similarity (section 3.1). Because the pair encoder evaluates ordered interactions among all $K+1$ fiber members, computation scales as $O(NK^2)$ per section A.3, so K directly trades off fiber informativeness against training and inference cost. We evaluate MolFiber-G, MolFiber-L, and MolFiber-ENS on the six OGB datasets at $K \in \{1, 8, 16, 24\}$.

Table 12: MolFiber-G, MolFiber-L, and MolFiber-ENS on OGB datasets across varying numbers of fibers K.

K	Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21	Avg. AUC ↑
1	MolFiber-G	85.42±0.00	71.21±0.30	98.55±0.30	80.22±0.30	66.56±0.40	76.93±0.10	79.81
	MolFiber-L	85.62±0.10	68.56±0.60	94.50±2.00	79.38±0.30	64.62±0.10	74.85±0.10	77.92
	MolFiber-ENS	85.47±0.00	70.80±0.30	98.62±0.50	80.27±0.30	66.46±0.80	76.90±0.40	79.75
8	MolFiber-G	85.45±0.00	72.53±0.70	99.17±0.30	80.04±0.10	66.25±0.30	77.47±0.30	80.15
	MolFiber-L	85.50±0.20	69.84±1.20	99.31±0.10	79.53±0.40	64.90±0.20	75.50±0.20	79.10
	MolFiber-ENS	85.49±0.10	72.94±0.70	99.29±0.10	80.33±0.50	65.80±0.20	77.09±0.30	**80.16**
16	MolFiber-G	85.46±0.00	72.13±0.90	99.25±0.20	80.35±0.30	66.32±0.20	77.56±0.20	80.18
	MolFiber-L	85.41±0.10	69.57±0.70	99.67±0.20	79.67±0.20	64.99±0.40	75.35±0.10	79.11
	MolFiber-ENS	85.42±0.00	72.47±0.80	99.62±0.20	80.48±0.40	65.98±0.40	77.31±0.30	**80.21**
24	MolFiber-G	85.49±0.10	72.62±1.30	99.14±0.50	80.16±0.70	66.31±0.20	77.28±0.20	80.17
	MolFiber-L	85.11±0.40	69.20±0.90	99.51±0.50	79.96±0.20	65.20±0.30	75.32±0.20	79.05
	MolFiber-ENS	85.29±0.20	72.31±1.40	99.53±0.40	80.43±0.30	65.98±0.30	76.94±0.20	80.08

Performance improves sharply from $K = 1$ to $K = 8$, a single neighbor gives too little context for the auxiliary views to resolve local ambiguity, costing MolFiber-L in particular 1.18 average AUC relative to $K = 8$, and continues to improve marginally from $K = 8$ to $K = 16$ (+0.03 for MolFiber-G, +0.01 for MolFiber-L, +0.05 for MolFiber-ENS). Beyond $K = 16$, performance plateaus or regresses: at $K = 24$, MolFiber-L and MolFiber-ENS both fall below their $K = 8$ values (-0.05 and -0.08 average AUC, respectively), indicating that sufficiently large fibers begin including Morgan-similar but less-informative neighbors that dilute rather than sharpen the within-fiber contrast.

We use $K = 8$ throughout the paper. While $K = 16$ yields marginally higher average AUC for all three methods, the improvement is small (at most $+0.05$ AUC) relative to the $4\times$ increase in pairwise fiber computation implied by the $O(NK^2)$ scaling ($K^2 = 64$ at $K = 8$ vs. $K^2 = 256$ at $K = 16$), and $K = 24$ ($K^2 = 576$, a $9\times$ increase over $K = 8$) yields no consistent improvement at all. We therefore select $K = 8$ as the setting that best balances performance against training and inference cost.

E FULL PER-DATASET RESULTS

This section reports per-dataset results, with standard deviations over $n = 5$ runs, for all thirteen baselines and MolFiber-G, MolFiber-L, and MolFiber-ENS on each of the 24 benchmarks, underlying the aggregate statistics in section 4. Table 13 repeats the summary table from main text for ease of reference alongside the full results below.

Table 13: **Molecular property prediction results (ROC-AUC).** Repeated from Table 1 for ease of reference alongside the full per-dataset results in this section. Comparison of 13 representative baselines from Praski et al. (2025) across Overall, OGB, and TDC benchmarks against MOLFIBER-G, MOLFIBER-L, and MOLFIBER-ENS. **Best, second, and third values are marked.**

	Overall (24 datasets)			OGB (6 datasets)			TDC (18 datasets)		
Method	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓	AUC ↑	Rank ↓	Gap ↓
CLAMP	**82.01**	**5.63**	**3.61**	**79.18**	6.42	**5.19**	**82.96**	**5.36**	**3.08**
R-MAT	**80.97**	6.46	4.65	77.75	7.58	6.62	82.04	6.08	3.99
CDDD	80.82	8.17	4.80	78.43	8.00	5.94	81.62	8.22	4.42
MolBERT	80.67	7.63	4.95	76.95	9.67	7.42	81.91	6.94	4.13
MAT	80.38	8.48	5.24	77.60	9.00	6.77	81.31	8.31	4.73
ChemBERTa	80.15	7.92	5.47	75.90	7.83	8.47	81.57	7.94	4.47
ECFP	80.14	7.79	5.48	76.67	6.83	7.70	81.29	8.11	4.74
MolFormer	79.80	9.90	5.83	77.20	9.75	7.17	80.66	9.94	5.38
Mol2Vec	79.59	10.35	6.03	78.12	7.42	6.25	80.08	11.33	5.96
Atom Pair	79.53	8.21	6.09	76.18	**5.08**	8.19	80.64	9.25	5.39
Uni-Mol	78.29	12.08	7.33	73.65	13.58	10.72	79.84	11.58	6.20
GraphMVP	77.65	12.58	7.98	73.82	12.67	10.55	78.92	12.56	7.12
GROVER	76.13	12.94	9.49	70.87	12.00	13.50	77.88	13.25	8.16
MolFiber-G	**82.01**	**5.19**	**3.62**	**80.15**	**6.33**	**4.22**	**82.62**	**4.81**	**3.41**
MolFiber-L	80.96	7.92	4.66	79.10	8.17	5.27	81.58	7.83	4.46
MolFiber-ENS	**82.10**	**4.77**	**3.52**	**80.16**	**5.67**	**4.21**	**82.75**	**4.47**	**3.29**

E.1 OGB RESULTS

Table 14 shows a sharply bimodal pattern across the six OGB datasets. On ClinTox, MolFiber variants take the entire podium, MolFiber-L best (99.31), MolFiber-ENS second (99.29), MolFiber-G third (99.17), a clean sweep against all thirteen baselines, including strong pretrained models such as MolFormer (88.40) and CDDD (93.60). On the remaining five datasets, no MolFiber variant places in the top three, though none trails badly either: on BACE, all three variants cluster around 85.4–85.5, within 1.6 points of the best baseline (AtomPair, 87.10) and ahead of ten of the thirteen baselines; on BBBP, MolFiber-ENS (72.94) sits within 0.16 of third place (GROVER, 73.10). The clearest gap is on SIDER and TOX21, where MolFiber trails the podium by 2.3–2.8 and 0.8–1.2 points respectively; both are consistent with the domain-overlap caveat noted in section 4 for CLAMP, which wins outright on both datasets and is contrastively pretrained on ChEMBL bioactivity data closely related to these endpoints.

A second, less visible pattern is variance. Several strong baselines show substantial run-to-run instability on the smaller OGB datasets, GROVER's ClinTox standard deviation is $\pm$16.40, MolFormer's is $\pm$15.90, and Uni-Mol's is $\pm$13.50, each larger than the entire gap between the best and worst method on that column. MolFiber variants, by contrast, are tightly concentrated throughout: their ClinTox standard deviations range from $\pm$0.10 to $\pm$0.30, and no MolFiber variant exceeds $\pm$0.80 standard deviation on any OGB dataset. This stability is not incidental to the ClinTox sweep above: several baselines' point estimates on ClinTox are within a few points of MolFiber's, but their variance is an order of magnitude larger, making MolFiber's advantage there far more reliable than the raw numbers alone would suggest.

E.2 TDC RESULTS

Across the eighteen TDC datasets, MolFiber variants place in the top three on eleven, including two further full podium sweeps: PAMPA_NCATS (MolFiber-ENS best at 80.86, MolFiber-L second, MolFiber-G third) and hERG_Karim (MolFiber-G and MolFiber-ENS tied best at 89.02, MolFiber-L second). On the CYP-enzyme-inhibition family specifically (CYP1A2_Veith, CYP2C19_Veith, CYP2C9_Veith, CYP3A4_Veith), MolFiber variants consistently take second and third place behind

Table 14: **Full results table for six OGB datasets.** Results are reported for mean AUC ± standard deviation over 5 runs.

Method	BACE	BBBP	ClinTox	HIV	SIDER	TOX21
CLAMP	83.40±1.30	70.10±2.10	76.10±3.90	**93.10±2.00**	**69.20±2.70**	**83.20±4.30**
R-MAT	83.00±2.60	72.70±1.00	86.40±6.40	77.70±0.10	**68.70±1.20**	78.00±3.10
CDDD	81.10±3.50	72.30±1.90	93.60±12.60	**81.40±3.50**	63.90±2.60	**78.30±3.40**
MolBERT	81.90±1.70	**74.30±2.70**	87.20±4.80	77.90±1.30	64.00±2.20	76.40±3.00
MAT	83.90±1.20	72.60±2.50	89.00±1.00	77.60±0.20	65.50±2.20	77.00±4.10
ChemBERTa	82.60±4.50	72.10±1.00	73.90±1.40	78.70±2.50	**69.20±1.90**	**78.90±3.50**
ECFP	**86.00±4.00**	72.00±4.90	76.30±7.80	78.90±3.70	**68.60±3.90**	78.20±4.70
MolFormer	83.70±6.80	71.50±2.70	88.40±15.90	78.00±1.30	66.60±1.70	75.00±1.80
Mol2Vec	82.30±1.20	73.00±3.20	89.30±12.10	78.00±4.40	68.30±2.40	77.80±3.40
AtomPair	**87.10±1.40**	**73.80±3.30**	68.50±3.10	**81.10±4.20**	**68.70±2.30**	77.90±3.60
Uni-Mol	82.60±3.40	66.10±3.10	82.40±13.50	76.10±0.10	62.20±2.60	72.50±5.80
GraphMVP	81.80±5.60	68.70±3.80	78.90±2.40	70.20±0.20	67.30±2.00	76.00±3.60
GROVER	85.40±4.60	**73.10±5.80**	66.90±16.40	73.00±5.50	57.90±3.10	68.90±4.90
MolFiber-G	85.45±0.00	72.53±0.70	**99.17±0.30**	80.04±0.10	66.25±0.30	77.47±0.30
MolFiber-L	**85.50±0.00**	69.84±1.20	**99.31±0.10**	79.53±0.40	64.90±0.20	75.50±0.20
MolFiber-ENS	85.49±0.10	72.94±0.70	**99.29±0.10**	80.33±0.50	65.80±0.20	77.09±0.30

CLAMP, which wins each of these outright, again consistent with the ChEMBL bioactivity-pretraining overlap noted for OGB's SIDER and TOX21, since enzyme-inhibition assays are precisely the kind of bioactivity endpoint that pretraining would be expected to favor. That CLAMP's advantage is concentrated on TDC rather than OGB is also consistent with its pretraining corpus: it is trained by supervised contrastive learning on ChEMBL bioassay records, pairing molecules with the textual descriptions of assays in which they are active, so its embedding carries a bioactivity-specific inductive bias that overlaps in domain with the ADME and toxicity endpoints making up much of the TDC collection, and that the OGB tasks reward less directly. MolFiber's weakest results cluster on the three substrate-classification datasets (CYP2C9_Substrate, CYP2D6_Substrate, CYP3A4_Substrate) and on SARSCoV2_3CLPro_Diamond and hERG, none of which see a MolFiber variant place on the podium; hERG in particular is dominated by an outlier GROVER result (95.30), more than five points above every other method.

The stability pattern observed on OGB recurs here, often more sharply. On HIA_Hou, CDDD's standard deviation is ±17.10, larger than the entire spread between best and worst method on that dataset, while MolFiber-G and MolFiber-ENS both stay under ±0.30. ECFP's DILI standard deviation (±10.10) and GROVER's SARSCoV2_3CLPro standard deviation (±12.80) show the same pattern: several baselines exhibit double-digit run-to-run variation on individual TDC datasets, most of which are small (under 700 molecules for the three substrate tasks, under 900 for DILI and SARSCoV2_3CLPro), while every MolFiber variant remains below ±1.4 standard deviation across all eighteen datasets. This consistency across both dataset pools is what the aggregate Av.AUC, Av.Rank, and Av.Gap statistics in table 1 are intended to capture: not merely competitive mean performance, but performance that is reliable enough, run over run, for the ranking itself to be trustworthy.

Table 15: **Full results table for TDC datasets.** Results are reported for mean AUC ± standard deviation over 5 runs.

Method	AMES	Bioavailability_Ma	CYP1A2_Veith	CYP2C19_Veith	CYP2C9_Substrate	CYP2C9_Veith
CLAMP	84.60±1.90	64.20±3.00	**95.10±1.30**	**92.20±2.40**	**69.80±3.50**	**91.60±1.90**
R-MAT	**85.50±1.20**	72.80±4.50	93.10±1.40	88.10±1.40	66.60±1.90	87.90±1.60
MAT	81.80±3.10	68.70±4.10	92.70±2.10	87.50±2.30	62.50±2.80	87.60±1.90
MolBERT	**85.40±2.00**	68.40±1.00	92.90±0.70	88.50±1.90	64.70±0.70	88.50±1.50
ChemBERTa	84.30±1.70	**74.50±4.60**	92.60±1.00	87.10±1.20	**69.20±7.10**	88.00±1.20
MolFormer	81.70±1.50	72.20±4.40	91.40±0.40	85.50±1.00	62.50±1.80	86.60±1.00
CDDD	83.60±0.90	66.90±1.20	92.40±0.90	87.20±0.30	65.50±1.90	88.10±0.60
ECFP	84.80±6.60	69.10±3.60	92.80±2.70	87.40±3.90	62.70±2.50	88.20±4.60
Atom Pair	81.50±3.20	71.20±2.00	92.90±1.90	85.80±2.00	59.60±2.60	87.20±2.20
Mol2Vec	80.60±2.10	72.60±5.00	91.30±0.90	85.20±2.10	64.50±4.40	86.40±1.10
Uni-Mol	80.60±3.20	**76.50±2.00**	91.60±2.10	86.30±3.20	**67.90±3.80**	86.50±2.60
GROVER	81.50±5.90	**88.40±1.30**	85.70±3.40	77.90±5.50	57.20±5.70	85.90±3.50
GraphMVP	80.80±3.70	64.90±4.00	88.90±1.10	85.00±1.70	67.40±1.10	83.80±1.60
MolFiber-G	**84.89±0.10**	71.27±0.60	**94.00±0.10**	**89.18±0.20**	64.53±0.90	**88.94±0.10**
MolFiber-L	84.32±0.10	71.27±0.60	93.63±0.10	87.99±0.20	62.71±1.10	87.66±0.20
MolFiber-ENS	84.87±0.10	71.34±0.60	**94.15±0.10**	**89.16±0.10**	64.72±1.40	**89.02±0.10**

Method	CYP2D6_Substrate	CYP2D6_Veith	CYP3A4_Substrate	CYP3A4_Veith	DILI	HIA_Hou
CLAMP	**84.30±1.50**	**89.30±2.50**	64.40±2.20	**90.40±1.80**	88.90±1.30	97.30±1.40
R-MAT	81.60±2.40	87.50±2.80	64.60±2.70	88.20±2.20	91.30±4.90	**98.70±1.50**
MAT	**83.00±1.30**	86.00±2.40	**68.10±5.30**	87.70±2.70	**93.50±2.70**	97.90±1.30
MolBERT	**82.20±2.50**	86.80±2.30	67.70±2.60	86.60±1.50	92.20±3.70	97.30±1.70
ChemBERTa	81.50±0.90	86.00±1.30	65.80±3.80	86.20±0.80	89.70±1.60	96.40±1.20
MolFormer	**83.00±2.10**	85.50±1.30	63.10±3.10	85.00±1.00	92.00±3.80	98.10±2.00
CDDD	81.00±1.80	86.80±2.40	65.90±0.90	86.30±0.80	92.30±3.00	94.30±17.10
ECFP	80.50±2.50	87.20±5.00	66.40±1.80	87.50±4.40	91.50±10.10	95.00±1.90
Atom Pair	79.40±0.50	85.80±4.40	**68.40±1.70**	86.70±2.20	92.90±4.70	**98.70±1.90**
Mol2Vec	81.20±1.00	84.20±3.70	63.50±0.50	84.40±0.70	92.20±7.90	**99.30±4.30**
Uni-Mol	79.60±3.90	85.40±4.20	60.60±2.80	86.10±2.50	92.10±1.60	94.90±4.00
GROVER	75.20±8.00	81.90±5.60	57.50±1.90	85.00±5.70	89.60±0.90	95.30±2.40
GraphMVP	79.30±1.10	82.80±2.20	**67.80±6.40**	83.10±1.90	90.20±4.20	93.10±5.10
MolFiber-G	80.93±0.80	**88.47±0.20**	64.44±0.70	**89.08±0.10**	**93.76±0.50**	98.17±0.20
MolFiber-L	78.27±0.50	87.74±0.10	63.93±0.10	87.30±0.10	92.53±0.20	96.59±0.60
MolFiber-ENS	80.65±0.40	**88.52±0.20**	64.72±0.70	**88.82±0.20**	**93.58±0.40**	**98.11±0.30**

Method	PAMPA_NCATS	Pgp_Broccatelli	SARSCoV2_3CLPro	SARSCoV2_Vitro	hERG	hERG_Karim
CLAMP	74.80±2.20	**93.30±0.80**	73.80±7.30	**63.20±4.20**	**90.00±1.00**	86.00±1.70
R-MAT	72.60±2.90	92.70±0.80	**76.60±2.70**	58.30±1.00	83.70±3.30	87.00±1.60
MAT	74.60±7.30	92.20±0.70	71.60±0.50	60.80±3.40	82.10±4.70	85.30±1.60
MolBERT	74.00±2.90	91.90±2.20	72.60±0.90	61.30±4.70	85.70±3.40	87.60±2.20
ChemBERTa	73.00±6.50	92.90±0.90	**77.10±3.60**	53.00±2.00	84.20±3.30	86.80±2.40
MolFormer	71.60±8.10	92.90±0.70	74.90±3.80	58.70±4.10	83.10±3.30	84.10±2.60
CDDD	76.20±6.10	91.80±1.00	74.00±4.40	62.80±8.60	**87.70±4.40**	86.30±3.00
ECFP	76.40±5.20	92.80±3.70	73.30±4.00	59.40±3.80	80.50±3.60	**87.80±3.70**
Atom Pair	70.50±2.40	90.30±0.30	72.70±2.60	56.50±5.00	84.40±2.40	87.10±3.30
Mol2Vec	72.00±5.20	89.40±1.80	72.50±2.20	54.50±5.20	82.30±1.10	85.40±2.50
Uni-Mol	73.60±7.40	87.80±3.10	73.50±2.60	55.30±2.00	78.10±3.10	80.70±3.40
GROVER	66.70±6.30	86.60±3.20	74.00±12.80	48.60±4.20	**95.30±2.00**	69.60±3.10
GraphMVP	69.30±2.00	87.70±1.50	**75.20±3.70**	55.20±1.70	81.50±2.30	84.60±3.20
MolFiber-G	**80.50±0.40**	**93.18±0.30**	74.63±0.40	**63.40±4.40**	78.84±0.70	**89.02±0.30**
MolFiber-L	**80.68±0.40**	92.18±0.10	74.36±0.10	62.59±3.50	79.39±0.70	**87.92±0.20**
MolFiber-ENS	**80.86±0.70**	**93.20±0.30**	74.41±0.10	**65.05±3.10**	79.22±0.60	**89.02±0.30**